ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.13, No12, April 2017
E-mail: jslt@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

基于Python专用型网络爬虫的设计及实现

贾棋然

(郑州成功财经学院,河南 郑州 451200)

摘要:网络爬虫一种网络机器人,也有人说是网页的蜘蛛。随着科技在生活和工作中的应用,计算机也成了人们最为依赖的工具,随着互联网的信息管理量的逐渐增加,搜索引擎也是这个时期被创造并投入使用。但是初被使用的搜索引擎是无法精确搜索到人们需要的信息,面对人们越来越多样的需求,这样的搜索引擎已经无法满足人们的需求了。所以就有人研究了一种专用性的网络爬虫,它能解决传统搜索引擎的出现的局限性,所以该文将会对Python专用型的网络爬虫进行分析和探讨。

关键词:网络爬虫;Python;数据的挖掘;搜索引擎

中图分类号:TP393 文献标识码:A 文章编号:1009-3044(2017)12-0047-03

DOI:10.14004/j.cnki.ckt.2017.1417

在很多用户进行搜索引擎的使用中,往往会出现很多不需要的信息,这就是传统搜索引擎的局限性。通过传统的搜索引擎进行信息的搜索中,还需要用户对搜索到的信息进行分析,最终寻找到自己需要的信息。就目前的网络发达现状,这样的搜索引擎是非常浪费时间的,而且准确性也不高,用户很容易丧失搜索的心情。所以,本文将会针对这一问题,对专用型的网络爬虫进行分析,提高信息检索的效率。

1 分析Python和爬虫系统设计需求

1.1 Python的网络爬虫

网络爬虫主要是通过每个网页的链接地址进行相关内容的查找,然后将结果直接传送给用户,不用通过人工进行浏览器的操作来获取信息了。而Python是一种广泛应用的脚本语言,它自身带有urllib2、urllib相关的爬虫基础库等,在Python语言的基础上开发出的一种开源软件则是Scrapy,它可以在Linux、Windows等多种操作系统中使用。如果被获取的网页经过大量的HTML源代码进行编写,这种情况下需要下载很多内容,但是用户可以在Scrapy爬虫系统上制定一部分模块,从而实现爬虫的功能。

1.2 爬虫系统设计需求

在进行网络爬虫系统的开发时,对系统建设进行分析是基础性问题,同时也要将符合设计该系统的代码和功能规范提出来。这样能够促进网络爬虫系统顺利的开发,进而保证开发的结果能够符合系统功能的基本需求。网络爬虫系统的建设基本上同时通过模块化进行的设计,一般每个功能都要自己的模块。这样能够方便以后进行代码的维护,而且还能提高代码的重要性。将整个系统分成不同的模块,之后把每个模块的功能编制完成,这样整个网络爬虫体系的功能就是已经完成了。本系统主要是根据某些用户的上网习惯,进行网络专用型的爬虫系统设计,根据用户的不同需求,确定网络爬虫系统中的各个功能。而且在进行系统的设计时,还要考虑系统以后的改进和维护等问题。本文会结合网易和豆瓣网等进行爬虫系统的分析,以及建立爬虫系统的内容。

1.3 功能需求分析

网易新闻关于爬虫的建立,主要包含几个功能部分:新闻的标题、新闻的来源、新闻的ID等,然后在将抓取的信息储存在数据库中。在网易新闻中需要爬虫获取的URL连接是动态的,并不是固定的,所以在建立爬虫的URL模块时,要先解决URL连接的去重,以及访问对策等问题。根据网易新闻中原站点的各种新闻状态,可以看出如果所有的新闻被发布完之后,就不会进行第二次的更新,所以在建立网易新闻爬虫的时候,当抓取结果的之后,就不必在对数据库中的信息进行更新了,只需要把网站上更新的内容储存到数据库中就行了。相关的爬虫结构框架可见图1。

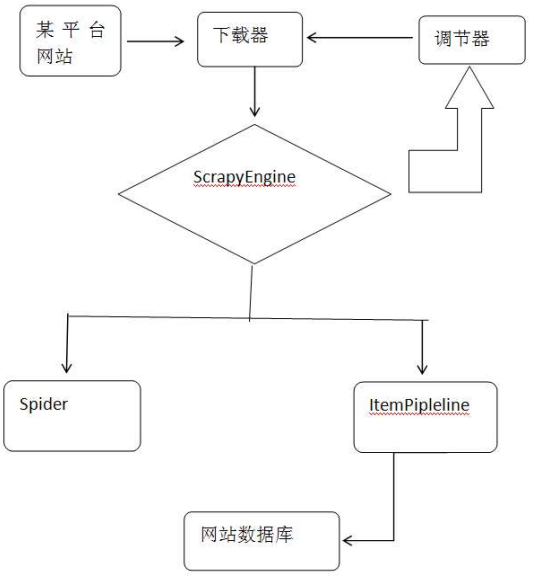

图1 网易新闻爬虫框架

1.4 爬虫功能的设计

网易新闻在进行爬虫功能的设计时,相关的逻辑是:首先,在网易新闻建设爬虫的时候,是不会将前端的页面和它之间进

收稿日期:2017-04-10

行交互的，网易是在系统中先进行任务的定时设置，之后爬虫根据定时的任务进行运作，如此实现了一种自动定时到网站原点进行新闻的抓取功能。其次，当网易爬虫被定时的任务驱使运作时，它会根据原有的URL规则，在原点站内相关的节点目录进行分析，将对符合需求的URL连接实施抓取工作，之后再将抓取的结果进行信息的提取和过滤。最后，把获取的新闻信息与数据库中的内容进行比较，如果数据库中没有这一条新闻的话，将新的数据插入到数据库中，如果有这样的信息就停止爬虫。该运行流程可以见图2。

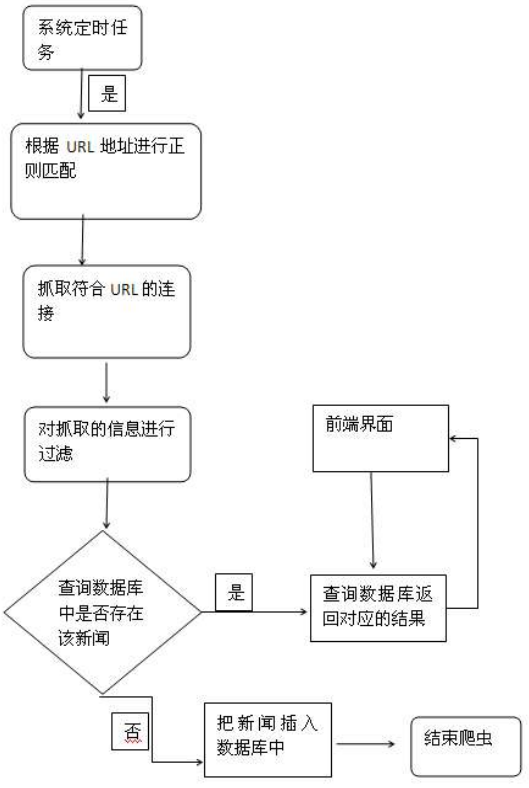

图2 网易新闻爬虫的流程

2 爬虫页面管理功能的分析

2.1 爬虫页面的抓取

进行爬虫页面的抓取任务是实现爬虫程序的第一步。在进行页面的抓取时，需要获取目标站点相关的动态，还要建立DNS解析与去重记录等功能。另外，在利用爬虫抓取页面的时候，也要保证抓取的目标站点是处于可抓取的状态。也就是说如果一些目标站只有用户登录之后，才能对服务器中的资源进行相应的请求，对于这样的站点就需要建立一个登录目标站点的模拟，在进行页面的抓取。模拟登录顺利通过目标站点有关登录的限制。这个方法主要是按照目标站点的规则进行的，它主要是利用用户名和密码、cookies和伪造的User-Agent与Referer等进入站点，然后将返回的session和服务器实现请求交互，之后在实施页面的抓取，进而完成整个抓取页面的任务。

在进行页面的抓取中，DNS解析以及去重URL记录环节是整个抓取模块的重要部分。在进行大量的页面抓取工作时，需要通过URL地址才能进行，所以在进行URL的请求时，必须先对URL进行解析。如果在进行URL解析的数据不较多的话，那么DNS的解析会成为抓取页面的瓶颈，想要解决DNS解析这一问题，可以对DNS解析的结果直接进行本地缓存。去重记录主要是对抓取完成之后的URL地址实施记录的去重。抓取页面的时间是有一定时间限制的，所以只能抓取一次。在进行页面的抓取时，要做好相应的记录去重，这样是为了避免出现重复抓取的情况，如此就会影响系统运行的性能和信息的高效性。

2.2 爬虫页面的处理

在抓取页面完成之后，还需要对页面进行一定的整理。进行页面处理时，首先需要对HTML相关的源代码进行过滤和处理，分析出需要的信息，然后在对分析出的结果信息进行整合，最后实施入库的操作。一般对页面进行处理时都是使用正则表达式，如果HTML源码比较多的话，在通过正则表达式进行编写时是比较困难的。

对HTML源码进行过滤时可以利用XPath操作，在进行不同的需求处理时，也要进行不同的XPath语法定义。比如：XPath语法中的get_title方法的使用：title=responsexpath（"/html/head/title/text()"）.extract()。也就是说值利用XPath语法，就可以对新闻标题进行过滤，也不必进行复杂的正则表达式编写了。比如：XPath语法中的get_source方法的使用：source=response.xpath("//div [@class='ep-time-sourec-DGray'/text()]").extract().经过这些都方法的使用，最后可以在页面处理模块得到一个有关新闻中某个新闻编号，或是新闻标题、新闻来源等信息。在获取这些原始数据之后，就将他们整合成一个列表，之后将其传递给爬虫入库。相关的流程可以见图3。

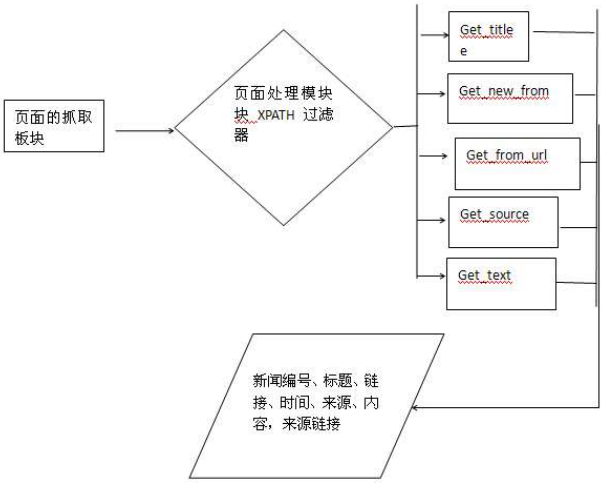

图3 网易新闻爬虫的页面处理

由于网页新闻要进行定期更新的，想要节省服务器中的资源利用，在网易新闻界面功能的实现中，可以将网易爬虫工作的时间设置为每小时自动更新一次。在新闻界面，将当天所有的新闻显示在上面，点击网易新闻的标题，就可以进入网站进行详细信息的了解。

3 基于Python专用型网络爬虫的实现

对基于Python实现网络爬虫的建设，这里主要以豆瓣网为例。首先要利用爬虫获得相关影视地址的连接，之后按照获得地址连接对目标影视信息进行解析，然后将其储存在对应的文件中。详细的过程如下：

3.1 获取需要爬取的影视URL

想要获得所有的影视URL，首先就是要找到URL相关的一个入口。比如：豆瓣网，它的入口网址是：https://www.douban.

com/tag/****/movie?Start=****，进入网站可以看到有15部影视被成列在页面上，在第一个“****”处需要抓取的信息是最早上映的影视，第二个“****”处就是将第一个要展示的影视作品相关的序号填写上去（这些序号可以是0,15,30……就是从0每次递增15）。然后将序号进行改变，如此就会实现了下一页影视的展示。比如：用start=0代表第一页，用start=15代表第二页，以此类推，每次增加15.总之就是将网址入口处的两个取值稍微改变 一下，就可以得到很多不同的影视信息。

获取影视URL的爬虫类有两种分别是：spiederUtil、doubanSpider。其中使用doubanSpider的过程：首先要在main函数中建立一个对象为doubanSpider,通过这个对象利用getcontents函数将影视列表所有的URL源码抓取到，然后在利用spiderUtil这个类保存这些文件。之后通过readALL函数浏览所有的文件，然后在使用parseWeb函数，将这些源码解析成影视数据，然后通过spiderUtil类使用save方法，保存这些影视作品。

3.2 解析影视信息

(1) 修改items.py相关文件。该文件对应的是TutorialItem类，利用这个类引出scrapy.item中的Field和Item类，最后在给影视进行定义。

(2) 修改Pipelines.py相关文件，它是管道处理的文件。

(3) 然后在将movieSpider.py文件进行编写。在spiders文件中，它是爬虫的主体。

(4) 最后就是将配置settings.py文件进行修改。针对一些具有防爬虫机制的网站，设置一些属性。

4 结束语

总之，在Python自带的库中获取网页信息，之后利用正则表达提取信息，然后利用Python中的Scrapy软件实现Web信息的抓获。

参考文献：

[1] 姜杉彪, 黄凯林, 卢昱江,等. 基于Python的专业网络爬虫的设计与实现[J]. 企业科技与发展, 2016(8):17-19.
[2] 唐哲炜. 基于python的网络爬虫设计与实现[J]. 科研,2016,6(18):28.
[3] 钱程, 阳小兰, 朱福喜. 基于Python的网络爬虫技术[J]. 黑龙江科技信息, 2016(36):273.
[4] 张丹. 基于Python的新浪微博数据爬虫程序设计研究[J]. 科研, 2017(2):00031.

（上接第46页）

关注的计算机网络信息安全的防护措施提供一定意义的帮助。网络信息安全随着网络技术的不断更替以及网络技术的更广泛和深入的应用，未来的网络信息安全奖逐渐成为关系国家民生、国家大事的重要保障网问题，面对这样严峻的形势，如何促进网络信息安全的升级换代，如何构建我国网络信息安全的防护体系已经成为了时代发展和我国国家安全工作发展需要重点关注和研究的问题之一，高等教育上要分析考虑设置专门的网络信息安全的专业学科的建设，把问题的研究推向一个高度，以便为将来网络信息安全防护提供必要的理论支持。

参考文献：

[1] 刘发胜.浅议计算机网络系统中的信息安全风险与防护措施[J].电脑知识与技术,2014(9):1876-1877.
[2] 来羽,张华杰.基于无线网络的网络信息安全防护措施探究[J].煤炭技术,2013(10):234-235.
[3] 靳元良.计算机网络信息安全及防护措施研究[J].中国管理信息化,2015(20):152.
[4] 龙震岳,魏理豪,梁哲恒,艾解清. 计算机网络信息安全防护策略及评估算法探究[J].现代电子技术,2015(23):89-93.
[5] 杨雨锋.计算机网络与信息安全防范措施探究——评《计算机网络与信息安全技术》[J].新闻与写作,2016(9):125.