DOI:10.14083/j.issn.1001-4942.2018.01.030

基于 Scrapy 的农业网络数据爬取

李乔宇,尚明华,王富军,刘淑云

(山东省农业科学院科技信息研究所,山东 济南 250100)

摘要: 准确、及时、高效地获取农业数据是全产业链农业信息分析预警工作的前提和基础,是提升农业信息分析预警专业化和规范化水平的关键。本研究针对互联网中存在的大量农业信息数据,以玉米价格数据为例,设计数据抓取和规范化存储策略,首先基于 Scrapy 框架建立对网页的请求响应,分析网页布局后对关键信息进行循环抓取,并利用正则表达式将抓取的信息提取为格式化数据,然后将数据本地化存储为 Microsoft Excel 表格或存储至数据库中,最后利用 Echarts 将数据以可视化的方式在 Web 端展示,从而实现对农业网络数据的挖掘和利用。

关键词: Scrapy; 爬虫; 网络数据; 数据挖掘; 玉米价格

中图分类号: S126 **文献标识号:** A **文章编号:** 1001-4942(2018)01-0142-06

Data Crawling from Agricultural Internet Based on Scrapy

Li Qiaoyu, Shang Minghua, Wang Fujun, Liu Shuyun

(*Institute of Science and Technology Information, Shandong Academy of Agricultural Sciences, Jinan* 250100, *China*)

Abstract Accurate, timely and efficient access to agricultural data is the prerequisite and basis for analysis and early warning of agricultural informations in the whole industry chain. It is the key to enhancing the professionalization and standardization of agricultural information analysis and early warning. With the maize price as an example, the research focused on large amounts of agricultural informations on the Internet and developed data crawling and normalized storage strategies. Firstly, we created request & response to the web pages based on Scrapy framework, analyzed the web page layout and then crawled the key informations cyclically; the data were extracted into formatted data using regular expressions, and then were stored as the localized data in a Microsoft Excel spreadsheet or in a database. Finally, Echarts was used to visualize the data on the Web, and thus the mining and utilization of agricultural network data were realized.

Keywords Scrapy; Crawler; Network data; Data mining; Maize price

随着大数据技术的发展,农业大数据的开发和利用逐渐成为当前研究的热点。农业大数据来源于农业生产、农业经济、农业流通、农业科技等各个方面,来源广,类型多,结构复杂,具有潜在应用价值。数据来源不同,其获取技术不同,目前农业大数据获取主要包括:农业生产环境数据采集、生命信息智能感知、农田变量信息快速采集、农业遥感数据获取、农产品市场经济数据采集、农业网络数据抓取等[1]。在“互联网+农业”的发展形势下,农业网络数据已成为农业大数据的重要组成部分,但由于其数据格式复杂多样,不利于快速统计分析,多仅是对数据的粗略展示,因此,如何有效统一农业网络数据格式,进一步挖掘数据的深层价值,成为当前大数据技术研究的重点。

收稿日期: 2017-07-25; **修回日期:** 2017-11-22

基金项目: 山东省农业科学院青年科研基金项目(2016YQN47);山东省重大应用技术创新项目“基于物联网的设施蔬菜大数据平台研究与应用”;山东省农业科学院农业科技创新工程项目(CXGC2016B15)

作者简介: 李乔宇(1986—),男,硕士,助理研究员,主要研究方向为图像分析、农业信息化。E-mail:joray86@126.com

通讯作者: 刘淑云(1974—),女,博士,副研究员,主要研究方向为农业信息技术。E-mail:13969079359@126.com

农产品市场价格信息对于分析农产品市场行情变化，预测其价格走势，降低交易风险，增加收益，具有重要意义。目前，网络上的农产品价格数据，一般是由特定工作人员采集市场价格信息后通过移动终端上报各农业服务机构，再由农业服务机构发布到网上[2,3]，对农产品交易具有一定的指导意义。但由于各服务机构发布的数据格式不统一，不利于对相关数据的进一步挖掘分析，限制了其利用价值。利用爬虫技术从网络中以一定的规则采集数据，并统一格式存储，为进一步挖掘网络数据应用价值奠定了基础[4-7]。

网络爬虫(web crawler)也叫网络蜘蛛(web spider)，是实现自动浏览网页和网页数据抓取的计算机应用程序。Scrapy 是使用 Python 编写的爬虫应用框架程序，具有结构简单、使用方便的特点，用户借助 Scrapy 可以快速浏览下载网页信息，并根据需要保存关键数据为需要的数据格式。目前，Scrapy 被广泛应用于数据挖掘领域，已经发展成为数据挖掘研究领域重要的应用工具[8]。

玉米是我国重要的粮食作物，玉米价格是市场发展和供给平衡状态的直接反映，通过提取其市场价格信息，不仅能够直观展示玉米交易市场的发展态势，同时有利于为供给侧结构性改革提供数据支撑，为相关部门制定生产发展决策提供理论依据。本研究以网络上发布的玉米市场价格为例，基于 Scrapy 设计爬虫，从中国饲料行业信息网爬取玉米价格信息数据，并以 Microsoft Excel 表格的形式存储或存为数据库，以期为农业网络数据的进一步挖掘利用提供一种有效的数据提取方法。

1　基于 Scrapy 的爬虫设计

1.1　Scrapy 框架

网络爬虫是以一定的规则自动抓取互联网信息的程序或者脚本，需要面向不同的应用场合解决网络连接、爬取策略等问题[9-12]。Scrapy 爬虫框架可以帮助开发者快速开发爬虫，其基于 Twisted 异步网络库来处理网络通讯，能够实现并行、分布式爬取，提高了爬取效率。

Scrapy 爬虫框架的结构如图 1 所示，包括以下 5 个主要模块：

①Scrapy Engine：引擎，负责 Spiders(爬虫)、Item Pipeline(队列)、Downloader(下载器)、Scheduler(调度器)之间的信息通讯和数据传递；

②Scheduler：调度器，负责接受引擎发送过来的 Requests(请求)，并按照一定的规则放入队列中；

③Downloader：下载器，负责下载 Scrapy Engine 发送的所有 Requests，并将其获取到的 Responses(响应)交还给 Scrapy Engine，由引擎交给 Spiders 来处理；

④Spiders：负责处理所有 Responses，从中提取数据，获取 Item 字段需要的数据，并将需要跟进的 URL 提交给引擎，再次进入 Scheduler；

⑤Item Pipeline：负责处理 Spiders 中获取到的 Item，并进行处理，如去重、持久化存储等。

Scrapy 的工作流程是：Scrapy Engine 启动并控制爬虫运行，首先由 Spider 根据编写的爬虫策略控制 Scrapy Engine 向 Scheduler 发送请求(Requests)，Scheduler 将请求加入队列，依次向 Downloader 发送，Downloader 接收请求后将互联网信息下载到本地成为响应(Response)，传递给 Spiders 处理后形成 Items，由 Pipeline 保存或输出。

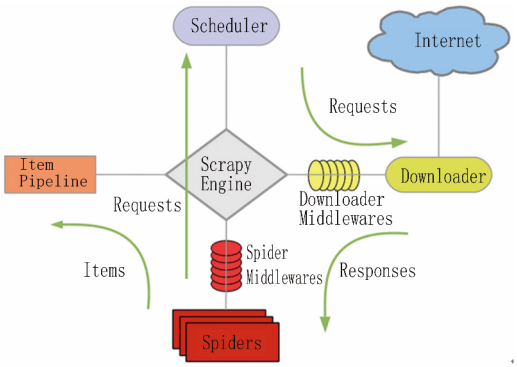

图 1　Scrapy 爬虫框架

1.2　玉米价格爬虫设计

1.2.1　玉米价格爬虫特点　网络爬虫分为全网爬虫和聚焦爬虫。全网爬虫面向整个互联网，目的是尽可能多地索引互联网资源[13]，是实现搜索引擎的主要工具；聚焦爬虫面向特定目标，目的是获取信息和数据，是信息挖掘的主要工具。玉米价格爬虫为聚焦爬虫，目标不是从网络资源中抓取特定网页，而是从特定网站分类和检索有价值的信息。因此，玉米价格爬虫不需要检索全网资源，负载小，设计重点是对价格数据的提取。

1.2.2　爬取策略设计　以中国饲料行业信息网

的玉米页面(http://www.feedtrade.com.cn/yumi/yumi_daily/)为例,见图2。矩形框①标注的区域为玉米价格的文章列表;椭圆框②标注的区域为待爬取的文章标题。玉米价格爬虫的目标,是在网页中定位文章列表区域,获取文章列表,从中筛选出希望爬取的文章标题,根据标题指向的链接进入文章内容页,从内容页中下载需要的信息到本地,不断循环直至获取所有需要的信息。

爬虫工作流程如图3,把入口点链接(http://www.feedtrade.com.cn/yumi/yumi_daily/)加入调度器(任务队列)中,调度器将任务分配给下载器,下载器将链接指向的页面下载到本地,根据解析规则判断该页面为文章列表页,从该页面获取文章列表,根据列表中的标题是否包含"山东"、"玉米"字段选择待爬取的文章标题,取出标题指向的内容链接放入调度器中,之后判断文章列表页是否有下一页,如果有就将下一页指向的链接放入调度器,依此循环,直至不存在下一页。这样就将所有的文章内容页链接放入到了调度器,当下载器下载到文章内容页,就会下载玉米价格文章,交由Pipeline做分析处理,提取玉米价格信息。当调度器中所有的链接都被下载,玉米价格信息爬取完成。

图2 待爬取的页面

2 爬虫实现

2.1 定义爬取对象

网页中包含许多内容,爬虫只抓取需要的内容。爬虫的Item定义了爬取的对象,在玉米价格爬虫中,用MaizeItem定义爬取对象,包括:文章标题、文章日期、文章链接、文章内容。

2.2 网页爬取

Web网页是结构化的,网页中的任何内容都处于结构体中,因此可以在结构体的路径中查询到。如图4所示,在Chrome浏览器中右键选择"检查",可以看到网页的结构体,框①显示网页中展示的玉米价格条目,框②为该条目在结构体中内容,其中包括文字信息和指向链接,框③为该条目在结构体中的路径。使用xpath()可以根据结构体的路径定位到玉米价格条目。在玉米价格爬虫中,程序写为:

```
1. content = response.xpath('//ul/li/a[contains(@title,"山东玉米")]')
```

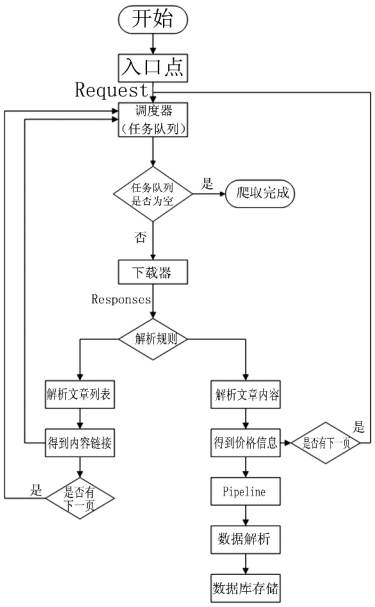

图 3　网页爬取流程

该段代码表示查找所有处于“ul/li/a”路径下，题名包含“山东玉米”的条目，从中提取文章标题、文章日期和文章链接放入 MaizeItem 中。文章链接存为 contentUrl，放入调度器中。对文章内容页的解析与文章列表页不同，使用 scrapy. Request(contentUrl， callback = self. parseMaizePrice) 指定解析规则是 parseMaizePrice( ) 。在 parseMaizePrice 中编写解析规则，同样使用 xpath 抓取文章内容，存入 MaizeItem 中，这样形成了一条包括文章标题、日期、链接、内容的完整 Item 条目，交给 Pipeline 处理。此外，在文章列表页和文章内容页使用 xpath( ) 定位“下一页”的链接，如果存在，则放入调度器中，继续循环。

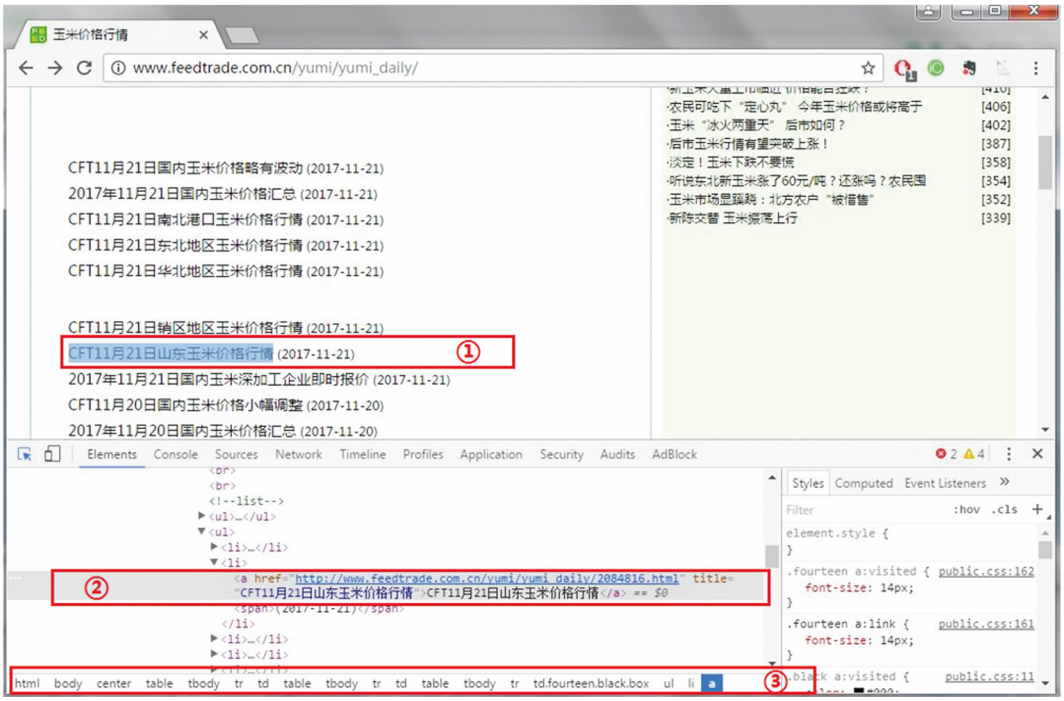

图 4　网页结构解析

2.3　数据提取

在 Pipeline 中对爬取的 MaizeItem 对象进行处理，形成按日期索引的玉米价格数据。其中，MaizeItem 中的文章内容是文字格式，形式如“……山东潍坊市寿光金玉米淀粉，价格统一调整为 1.6 元/千克。潍坊寿光新丰淀粉，水分 30 以内价格执行 1.61 元/千克，落 0.01 元……”，该内容信息分散，难以利用，需要对其进行进一步提取，以形成方便利用的价格数据。经分析，MaizeItem 数据中包含的多个公司的玉米价格信息，只有部分文字内容和价格数据是每日变动的，而公司名是每日重复的。因此，数据提取的任务是：忽略文字内容，提取出公司名称和价格数据，并建立二者的关联关系。

首先整理出公司名称列表，根据该列表，使用正则表达式查找并提取该公司的玉米价格。正则

表达式以“\d”表示数字，使用匹配规则“公司名称+‘.*？(\d+.\d*)元/千克’”可以查找到公司名称下以“元/千克”为单位的价格数字。但价格数据的类型并不完全相同，有的是“X.X元/千克”的形式，还有的是“XX元/吨”或者“XX-XX元/吨”的形式，所以需要对公司列表根据价格数据的类型进行分类，然后用不同的正则表达式提取价格数据。数据提取的结果是公司列表中公司每日的玉米价格数据。

2.4 数据存储

对于爬取得到的数据可以本地化存储，也可以存入数据库中。

2.4.1 本地化存储

虽然Pipeline中可以直接建立json文件将数据写入，但json文件可读性比较差，因此，本研究尝试将json文件进一步处理后保存为可读性强的excel文件。提取到的公司每日玉米价格数据表是一个二维表结构，Pandas库是python中处理数据的主要工具，借助其中的DataFrame可以存储二维表结构。首先建立｛公司名:玉米价格｝”的字典priceDict，再建立｛日期:priceDict｝”的二维字典priceAll，将其保存为DataFrame，然后以公司列表建立DataFrame的索引，最终使用DataFrame的to_excel()方法保存为Excel文件(图5)。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		2017-3-20	2017-3-21	2017-3-22	2017-3-23	2017-3-24	2017-3-25	2017-3-26	2017-3-27	2017-3-28	2017-3-29	2017-3-30	2017-3-31	2017-4-1	2017-4-2
2	潍坊市寿光金玉米淀粉	1.656	1.656	1.656	1.656	1.69	1.69	1.73	1.73	1.73	1.724	1.724	1.724	1.734	1.734
3	潍坊寿光新丰淀粉	1.66	1.66	1.66	1.66	1.69	1.69	1.73	1.73	1.724	1.724	1.72	1.72	1.73	1.73
4	潍坊市诸城兴贸淀粉	1.65	1.65	1.65	1.65	1.66	1.66	1.66	1.66	1.66	1.66	1.67	1.68	1.7	1.7
5	潍坊市昌乐盛泰	1.676	1.676	1.676	1.676	1.714	1.714	1.746	1.746	1.74	1.74	1.734	1.734	1.744	1.744
6	潍坊市昌乐英轩实业	1.69	1.69	1.69	1.69	1.71	1.71	1.71	1.73	1.73	1.73	1.76	1.76	1.76	1.76
7	聊城市临清金玉米	1.63	1.63	1.65	1.65	1.66	1.66	1.69	1.69	1.69	1.69	1.69	1.69	1.71	1.71
8	滨州市邹平西王	1.65	1.65	1.66	1.66	1.68	1.68	1.7	1.7	1.72	1.72	1.72	1.72	1.72	1.72
9	滨州金汇	1.62	1.62	1.62	1.62	1.644	1.644	1.644	1.66	1.66	1.66	1.66	1.66	1.67	1.67
10	德州市平原福洋生物	1.64	1.64	1.65	1.65	1.68	1.68	1.72	1.72	1.71	1.71	1.71	1.71	1.73	1.73
11	德州禹城保龄宝	1.64	1.64	1.65	1.65	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.73	1.73
12	德州地区粮点新玉米收购价格	1.53	1.53	1.53	1.53	1.575	1.575	1.575	1.575	1.615	1.615	1.625	1.625	1.625	1.625
13	临沂市沂水大地玉米	1.636	1.636	1.636	1.636	1.636	1.636	1.636	1.636	1.646	1.646	1.66	1.67	1.69	1.69
14	临沂市鲁洲集团山东公司	1.7	1.7	1.72	1.72	1.73	1.73	1.73	1.74	1.74	1.74	1.75	1.75	1.76	1.76
15	临沂地区饲料厂	1.64	1.64	1.64	1.64	1.64	1.66	1.66	1.66	1.68	1.68	1.68	1.68	1.7	1.7
16	山东泰安东平祥瑞药业	1.68	1.68	1.69	1.69	1.71	1.71	1.72	1.72	1.74	1.74	1.74	1.74	1.76	1.76
17	恒仁工贸有限公司	1.69	1.69	1.69	1.69	1.7	1.7	1.73	1.73	1.73	1.73	1.73	1.73	1.76	1.76
18	菏泽成武大地	1.63	1.63	1.626	1.626	1.636	1.636	1.636	1.666	1.666	1.666	1.666	1.666	1.7	1.7
19	青岛地区贸易商	1.57	1.57	1.57	1.57	1.59	1.59	1.59	1.59	1.62	1.64	1.64	1.65	1.65	1.65
20	青岛地区饲料企业	1.68	1.68	1.68	1.68	1.68	1.7	1.7	1.7	1.72	1.72	1.72	1.72	1.76	1.76
21	聊城地区贸易商	1.54	1.54	1.54	1.54	1.55	1.55	1.55	1.55	1.59	1.62	1.62	1.62	1.62	1.62
22	聊城地区饲料企业	1.64	1.64	1.64	1.64	1.64	1.64	1.64	1.64	1.68	1.68	1.68	1.68	1.71	1.71
23	聊城地区深加工企业	1.63	1.63	1.63	1.65	1.65	1.66	1.66	1.66	1.69	1.69	1.69	1.69	1.69	1.69
24	潍坊地区饲料厂	1.64	1.64	1.64	1.64	1.64	1.66	1.66	1.66	1.68	1.68	1.68	1.68	1.74	1.74
25	寿光地区深加工	16.64	16.64	16.64	16.64	0	1.69	1.69	1.69	1.724	1.724	1.722	1.722	1.722	1.722

图5 Excel格式的本地存储结果

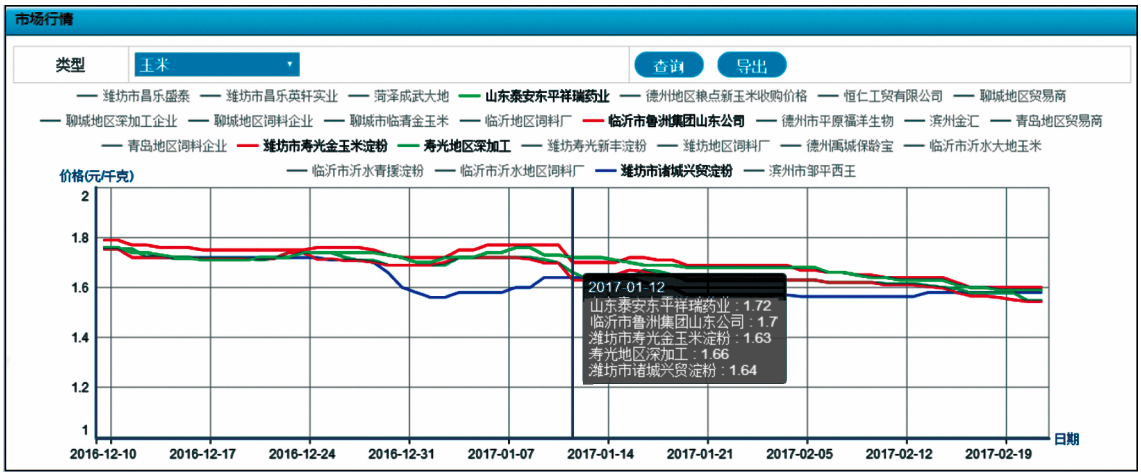

图6 数据爬取结果的Web展示

2.4.2 数据库存储

按照Scrapy中定义的数据模型在数据库中建立表，使用python的第三方模块pymysql登录数据库，执行sql语句向数据库中插入数据。为实现增量爬取，需要在数据表中建立校验字段，以内容链接的MD5加密信息作为校验信息存储，在插入数据前，先查询数据的校验信息是否已经在数据库中存在，防止插入重复的数据。

2.5 数据爬取结果展示

将爬取的玉米价格数据利用Echarts控件在

Web 中展示（如图 5）。项目共享在 Github 上：https://github.com/joray86/maizeSpider。

3　结语

本研究实现了基于 Scrapy 的农业网络数据爬取，爬虫运行环境为 CPU i5 - 4210U、4GB 内存、Windows 7 64 位操作系统，爬取中国饲料行业信息网（http://www.feedtrade.com.cn/yumi/yumi_daily/）的山东玉米价格信息，共访问页面 228 个，耗时 6 383 ms，提取到山东省的玉米市场流通价格数据，可在 Web 中进行可视化展示，也可持久化存储到数据库中或者存为 Excel 文件，这为用大数据方法进一步挖掘有价值的信息、市场信息监测和预警分析，以及政府决策和供需结构调整等提供了有力、可靠的数据支撑。

利用爬取的山东省玉米市场价格数据，可实现玉米市场价格的年度间同比和月、日数据环比分析，获得玉米市场价格的变化趋势，可为玉米产业的生产、流通、加工各环节起到一定的引领和指导作用。

参 考 文 献：

[1] 王文生，郭雷风. 农业大数据及其应用展望[J]. 江苏农业科学，2015，43(9)：43 - 46.

[2] 张石锐，郑文刚，申长军，等. 嵌入式手持无线农产品价格信息采集终端[J]. 计算机工程与设计，2012，33(2)：514 - 518.

[3] 尚明华，秦磊磊，王风云，等. 基于智能手机的小麦生产风险信息采集系统[J]. 农业工程学报，2011，27(5)：178 - 182.

[4] 段青玲，魏芳芳，张磊，等. 基于数据的农业网络信息自动采集与分类系统[J]. 农业工程学报，2016，32(12)：172 - 178.

[5] 孟繁疆，姬祥，袁琦，等. 农产品价格主题搜索引擎的研究与实现[J]. 东北农业大学学报，2016，47(9)：64 - 71.

[6] 郭雷风. 面向农业领域的大数据关键技术研究[D]. 北京：中国农业科学院，2016.

[7] 李慧，何永贤，叶云. 基于聚焦爬虫的农业信息服务平台设计与实现[J]. 天津农业科学，2016，10(10)：60 - 63.

[8] 马联帅. 基于 Scrapy 的分布式网络新闻抓取系统设计与实现[D]. 西安：西安电子科技大学，2015.

[9] 时永坤. 基于 Web Driver 的定向网络爬虫设计与实现[J]. 软件，2016，37(9)：94 - 97.

[10] 杜彬. 基于 Selenium 的定向网络爬虫设计与实现[J]. 金融科技时代，2016，7(7)：35 - 39.

[11] 赵本本，殷旭东，王伟. 基于 Scrapy 的 GitHub 数据爬虫[J]. 电子技术与软件工程，2016，6(6)：199 - 202.

[12] 夏火松，李保国. 基于 Python 的动态网页评价爬虫算法[J]. 软件工程，2016，19(2)：43 - 46.

[13] Castillo C. Effective web crawling[J]. ACM SIGIR Forum，2005，39(1)：55 - 56.