基于 Python 的新浪新闻爬虫系统的设计与实现

文/于韬　李伟　代丽伟

摘要

随着大数据时代的到来，数据量呈几何倍增长。以新浪新闻为代表的一系列新闻检索网站蕴含着大量的数据资源。本文以新浪新闻为研究对象，利用 Python 爬虫技术实现网页下载与网页解析，完成了对目标数据的高效获取，并将获取的信息进行格式化存储。实验结果表明，本文所提出的程序实现了网页数据的快速获取，为后续的数据挖掘提供支持。

【关键词】大数据 Python 爬虫 新浪新闻

1 引言

新浪新闻由新浪官方出品，及时发布全球新闻资讯，国内国外要闻，精彩的体育赛事报道，金融财经动向，影视娱乐事件，还有独家微博“微”新闻，精彩随你看，新闻、星座、笑话一个都不少。新闻是我们生活中的一部分，通过新浪的新闻板块可以坐在家里看世界。如此多的新闻信息，其中蕴含的巨大信息量是不言而喻的，因此如何获取是十分关键的。本文我们将通过爬虫技术获取相关新闻信息。

Python 作为一种语法简洁的程序设计语言，对于爬虫开发上有很多优势，在发送 HTTP 请求时，Python 提供优秀的第三方包譬如 Requests，极大简化了对网站的访问请求。在解析 HTML 源码时，提供的 BeautifulSoup 库能用极简短的代码完成过滤 html 标签并提取文本的工作。利用 Python 中的 pandas 可以对获取到的数据进行整理、储存。对于网站的反爬机制，Python 提供了更为简便的解决方案，可以使用 Requests 库得到一个代理 IP。Python 拥有足够多的简洁的语法和库的支持，使得它在爬虫开发上具有很高的效率。

本文提出的爬虫程序通过获取相关新闻信息，并将数据保存到本地，方便对数据的挖掘与分析。使用本程序可以节省获取数据的时间，使用户可以将更多精力放在数据分析上面。

2 基于Python的新浪新闻爬虫设计

2.1 爬虫系统设计需求

设计爬虫系统需要解决以下几个问题：

（1）评论数的获取：通过页面链接获取新闻 id，然后传递获取评论数。

（2）页面信息的提取：页面上有我们需要的标题、作者、摘要等信息。

（3）分页链接的获取：获取不同分页的链接，以便获取更多页面信息。

（4）网页内容的分析和管理：提取网页信息，并将其存入数据库或其他数据文件中。

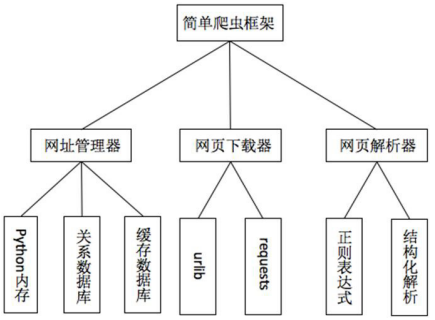

图 1：简单爬虫框架

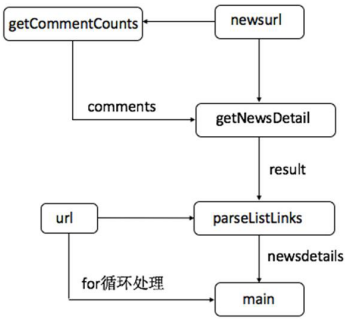

图 2：爬取流程图

	A	B	C	D	E	F	G
1		article	comments	editior	source	time	title
2	0	原标题：定了！公安边防、消防、警	5	责任编辑：桂强	人民网	2018年03月21日 16:57	定了 公安边防消防警卫部队全部退出现役
3	1	中新网3月21日电 据交通部网站消息	0	责任编辑：桂强	中国新闻网	2018年03月21日 16:52	出租车司机有酒驾吸毒记录等将被撤销从业资格证
4	2	原标题：4月9日起天津等6省市开展	0	责任编辑：张义凌	新京报	2018年03月21日 16:45	6省市将试点银行联网核查失效居民身份证信息
5	3	原标题：2018年高考进一步提高中西	0	责任编辑：桂强	新华网	2018年03月21日 16:42	2018年高考将提高中西部地区及人口大省录取率
6	4	原标题：一目了然！党政军地群团改	0	责任编辑：张义凌	新浪综合	2018年03月21日 16:42	一目了然 党政军地群团改革方案精简版60条
7	5	原标题：2018年世界杯期间中国观众	0	责任编辑：桂强	人民网	2018年03月21日 16:42	2018年世界杯期间中国观众可凭门票等免签入俄
8	6	原标题：央行新规：丢失身份证，再	8	责任编辑：桂强	中国新闻网	2018年03月21日 16:35	央行新规：丢身份证 再也不用担心名下多几百万
9	7	新华社快讯：中国驻马来西亚大使馆	0	责任编辑：桂强	新华国际	2018年03月21日 16:32	载16名中国船员挖沙船在马水域翻覆 致1死12失踪
10	8	原标题：中国理论海外传播研讨会在	0	责任编辑：桂强	人民日报海外版-海外网	2018年03月21日 16:30	中国理论海外传播研讨会在京召开
11	9	原标题：人物|郭树清出任银保监会主	0	责任编辑：张义凌	澎湃新闻	2018年03月21日 16:12	郭树清任银保监会主席：当证监会主席时7天一新政
12	10	原标题：新闻工作者援助项目实施办	0	责任编辑：张义凌	澎湃新闻	2018年03月21日 15:52	新闻工作者援助项目办法修订 因公殉职援助30万
13	11	定了！将组建中央广播电视总台 新	491	责任编辑：张义凌	人民日报	2018年03月21日 15:49	我国将组建中央广播电视总台
14	12	[独家]郭树清任中国银保监首任主席	1	责任编辑：张义凌	经济观察报	2018年03月21日 15:39	郭树清任中国银保监首任主席
15	13	原标题：[关注]北京新机场内景震撼	0	责任编辑：张义凌	政府网站	2018年03月20日 15:34	北京新机场内景亮相 犹如钢铁森林(图)
16	14	原标题：载16名中国船员的抽沙船大	1	责任编辑：桂强	中国新闻网	2018年03月21日 15:29	载16名中国船员抽沙船在马水域翻覆 致1死11失踪
17	15	原标题：定了！党中央机构将有这些	71	责任编辑：张义凌	人民日报	2018年03月21日 15:21	党中央机构将有这些重磅调整
18	16	原标题：我国等牵头制定的电动汽车	1	责任编辑：桂强	新华网	2018年03月21日 15:18	中国牵头制定的电动汽车安全全球技术法规获通过
19	17	原标题：吉林省人民政府：省长景俊	0	责任编辑：张玉	澎湃新闻	2018年03月21日 15:09	吉林省长景俊海兼任省地方志编委会主任
20	18	原标题：前国级卸任前的最后一次官	0	责任编辑：张玉	新浪综合	2018年03月21日 15:02	周小川卸任前参加最后一次官方活动 去了哪?
21	19	原标题：改革后，国务院首个总局局	0	责任编辑：张玉	新浪综合	2018年03月21日 15:02	改革后国务院首个总局局长亮相
22	20	原标题：就中美经济对话发矛盾信息	0	责任编辑：桂强	参考消息	2018年03月21日 14:45	外媒：就中美经济对话发矛盾信息 美国在“拍”啥
23	21	原标题：打击诈骗不分两岸 台湾警方	0	责任编辑：张岩	中国新闻网	2018年03月21日 14:23	打击诈骗不分两岸 台湾警方引用大陆影片进行宣传
24	22	原标题：中华遗嘱库：所保管8万余份	0	责任编辑：张玉	新京报	2018年03月21日 14:09	中华遗嘱库8万余份遗嘱中 超3成系老人瞒子女订立
25	23	原标题：保卫黄河：两“拦沙水库”	2	责任编辑：张玉	澎湃新闻	2018年03月21日 14:13	黄河两座“拦沙水库”是否修建引争论 正进行环评
26	24	原标题：生态环境部和环境保护部，	2	责任编辑：张岩	中国新闻网	2018年03月21日 14:15	生态环境部和环境保护部 两字之差有深意
27	25	原标题：广东水利厅原副厅长落马轨	11	责任编辑：张玉	澎湃新闻	2018年03月21日 14:13	广东水利厅原副厅长落马:4年送厅长80万 直接提拔
28	26	原标题：16部门：猴群服务机构保证中	2	责任编辑：张岩	中国新闻网	2018年03月21日 14:05	16部门：猴群服务机构保证中笃价位 严禁强制消费
29	27	原标题：台媒：对大陆“成急用忍”	0	责任编辑：张义凌	中国新闻网	2018年03月21日 14:08	台媒：民进党当局对大陆“成急用忍”是自丧机会
30	28	原标题：五角大楼称美国无力对抗高	27	责任编辑：张岩	环球网	2018年03月21日 13:50	五角大楼称美国无力对抗高超音速武器 又扯上中俄
31	29	原标题：教育部：中小学竞赛的获奖	0	责任编辑：张玉	新京报	2018年03月21日 13:29	教育部:中小学竞赛获奖结果不得作为招生入学依据

图 3：已抓取的新闻信息

2.2 与爬虫相关的python模块

2.2.1 网址管理器

实现网址管理的方法有以下 3 类：

（1）Python 内存储存：适合存储少量信息，将网址在储存时分为两类：已爬取和待爬取，放入两个集合中进行管理。

（2）关系数据库储存：适合网址信息进行永久性储存，可以存到表中，建立两个字段用来辨别是否爬取。

（3）缓存数据库储存：适合储存大量的网址信息

2.2.2 网页下载器

网页下载器是爬虫程序的主要核心模块。网页的内容一般是 HTML 格式，Python 支持的网页下载工具有两类：

（1）Python 官方支持的基础模块中的 urllib 包

（2）requests 第三方工具包，功能强大。

2.2.3 网页解析器

网页解析器是对网页内容中进行数据解析的工具。Python 支持的网页解析器有两种：一种利用正则表达式可以将整个网页文档当成一个字符串，使用模糊匹配的方式来提取出有价值的数据；另一种是根据 Html 网页创建成一个 DOM 树，以树的形式进行各种节点的搜索遍历。DOM 的树形结构根据上下级关系，可以很方便的定位到各个元素。

2.2.4 数据导出

利用 Python 可以将数据导出为 Excel 格式或其它格式文件。导出数据时应注意数据编码问题，否则导出的文件可能会出现乱码。如图 1 所示。

3 实验设计

我们通过新浪新闻 API：http://news.sina.com.cn/china/ 进行抓取。爬虫代码由 3 个功能函数和一个主函数构成：

3.1 获取详细页面内文函数（getNews

<< 下转 242 页

基于模块化的智能扩展云监控系统

文/郝佳琦　胡云生　谢雅丽

摘要

本文设计了一种基于模块化的智能扩展云监控系统，采用不同的传感器，实现了模块化的数据采集，针对不同的控制对象采用相应的传感器模块，实现该设备的智能云监控，用户可通过移动端实现与设备之间的通信，包括对设备工作信息的查看、对设备启停的控制和运行流程设定等。另外，本文对自动浇花系统和水位监控系统进行了系统测试，系统实现了在不同传感器模块情况下分别实现不同的控制流程及系统功能。

【关键词】智能家居 云监控 传感器

1 引言

智能家居不仅具有传统的居住功能，还兼备建筑、网络通信、信息家电、设备自动化，提供全方位的信息交互功能，甚至为各种能源费用节约资金，为生活提供了极大的便利。但目前大多数家庭所使用的家具设备大多属于非智能家居，可通过智能化改造将其升级为智能家居，如果对多数家居都进行智能化改造，一是难度大，二是成本高。本文设计了一种基于模块化的智能扩展云监控系统，可以在不同需求下通过连接不同的传感器模块将部分非智能设备升级为智能设备。消费者可以按需配置，实现不同家居的智能化改造。

表 1：时间 t 和湿度 p 的对应关系

时间 t/s	30	60	90	120	150
湿度 p01（实时）	300	344	396	420	514
湿度 p02（20min 后）	399	512	583	820	934

表 2：单片机参数和实际水深对应关系

实际水深 p0（mm）	0	40	80	120	160
单片机参数 p1（mm）	536	495	450	405	370
换算水位 P（mm）	4	45	90	135	170

2 系统设计与实现

2.1 整体设计思路

首先，设计单片机控制系统，使其能够完成对部分家具的控制，如控制电源、红外遥控等。其次，为了让控制系统实现远程控制，本文采用 Wi-Fi 模块通过串口与单片机连接，当 Wi-Fi 模块连接网络后可采用移动端与控制系统进行通信。最后，为了实现智能监控，需要为控制系统配备相应的数据采集功能，以返回家居设备工作状态，由此，可在移动端实现对家居设备状态的远程监控。

为满足不同情况下监视数据类型的不同，

<< 上接 188 页

Detail）

首先对详细页面的链接进行下载：通过 requests 方法下载 html 文档，接着通过 BeautifulSoup 进行解析。然后通过 select 方法获取文章标题、来源以及编辑信息，由于时间的格式的特殊性，利用 datatime 获取新闻发表的时间。对于新闻主体，通过 for 依次取段落，再通过 join 方法将所有段落信息整合在一起。取评论数我们通过 getCommentCounts 方法进行获取。最后定义一个字典，将页面的标题、来源、时间等信息存储到字典中。

3.2 获取评论数函数（getCommentCounts）

由于通过直接观察元素信息时找不到评论数信息，可能是通过 JS 方式添加上去的，因此需要对评论数链接进行处理，而评论数链接与新闻 id 有关，因此我们首先对新闻 id 进行获取。通过正则表达式获取新闻 id，在将新闻 id 放入评论数链接中，对此链接进行解析，即可得到相应的评论数。

3.3 剖析清单链接函数（parseListLinks）

在前两个函数中我们已经获取到了页面的详细信息，利用 parseListLinks 函数，我们获取不同清单的链接，并结合 getNewsDetail 函数获取清单上所有新闻信息。接着定义一个列表 newsdetails，并将 getNewsDetail 函数获取的信息储存在列表中。

3.4 主函数

在主函数中可以自定义想要获取的新闻页数。由于每个清单有许多数据页，我们加了一个 for 循环获得分页链接，通过 parseListLinks 方法依次获取新闻信息，我们利用 pandas 中 DataFrame 方法对数据进行整理，最后用 to_excel 方法将数据保存为 Excel 格式。

爬取流程图如图 2 所示。

4 实验结果

通过实验获取新浪新闻的论文信息，运行结果如图 3 所示，实验采集了 7500 条信息，为后期的数据处理提供了有力支撑。

5 结语

文章分析了新浪爬虫获取数据时的细节实现，对国内新闻 API 进行爬取，使用 requests 方法下载网页 html 文档，并用 BeautifulSoup 进行解析，进而获得相关的数据信息。总体来说，爬虫技术具有较高的应用价值与无限的潜在价值，通过抓取数据，可以挖掘出更有价值的信息。

参考文献

[1] 魏冬梅，何忠秀等，基于 Python 的 Web 信息获取方法研究 [J]. 软件导刊，2018(01).
[2] 孙立伟，何国辉等，网络爬虫技术的研究 [J]. 电脑知识与技术，2010(05).
[3] 周中华，张惠，然谢江，基于 Python 的新浪微博数据爬虫 [J]. 计算机应用 2014(11).
[4] 张明杰．基于网络爬虫技术的舆情数据采集系统设计与实现 [J]. 现代计算机（专业版）2015(06).
[5] 朱烨行，张明杰．微博数据采集的设计与实现 [J]. 电脑编程技巧与维护，2017(09).

作者单位

辽宁科技学院　辽宁省本溪市　117004