基于 WebMagic 爬取技术的电力事故信息获取

党 佩,阎光伟
(华北电力大学 控制与计算机工程学院,北京 102206)

摘 要:当前国民经济正处于迅猛发展的大好时期,也是电力工业体制改革的关键时期,对电力的需求十分紧迫,所以,电力系统的安全稳定运行及人员的安全管理日益成为影响电力工业发展的关键要素。近年来,各类电力事故依旧时有发生,全面调查事故发生原因是非常必要的,因此,进行事故信息的收集、管理和分析成为关键的一步。采用传统的方式,人工使用搜索引擎搜索信息,费时费力,而随着互联网技术的不断发展,网络爬虫技术已日渐成熟,应用网络爬虫技术可以快速获取这类事故信息。文中主要应用 WebMagic 爬虫技术,利用 XPath 和正则表达式指定信息的抽取规则,从电力安全管理网上抓取有关于电力事故信息的新闻,匹配符合要求的事故描述信息,下载到本地并实现数据存储进数据库,为之后进行事故信息分析提供数据基础。实验结果显示,该技术能够准确、迅速地获取数据,且爬虫程序简单易维护。

关键词:电力事故;网络爬虫;WebMagic;数据抓取
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2019)06-0125-05
doi:10.3969/j.issn.1673-629X.2019.06.026

Acquisition of Electric Power Accident Information Based on WebMagic Crawling Technology

DANG Pei, YAN Guang-wei
(School of Controland Computer Engineering, North China Electric Power University, Beijing 102206, China)

Abstract: At present, the national economy is in a great period of rapid development, which is also a crucial period for China's electric power industry system reform, and the demand for electric power is quite urgent. Therefore, the safe and stable operation of electric power system and the safety management of personnel are increasingly becoming the key factors affecting the development of the electric power industry. In recent years, various types of electric power accidents have occurred from time to time. It is necessary to investigate the causes of these accidents. Therefore, the collection, management and analysis of accident information has become a crucial step. In the traditional way, manually using search engines to search for information is time-consuming and laborious. With the continuous development of Internet technology, web crawling technology has become more and more mature, by which such accident information can be obtained quickly. We mainly use WebMagic crawler technology, and apply XPath and regular expressions to specify the information extraction rules, grabbing news about electric power accident information from the Electric Power Security Management Network, matching the accident description information that meets the requirements and realizing data storage into the database, which provides a data foundation for subsequent analysis of accident information. The experiment shows that the proposed technology can acquire data accurately and quickly, and the crawler program is simple and easy to maintain.

Key words: electric power accident; web crawler; WebMagic; data crawling

0 引 言

电力生产是把各种一次能源,包括化石燃料(煤炭、石油、天然气)、可再生能源(水能、风能、太阳能、潮汐能、地热能和生物质能等)以及核能转换成电能,并输送和分配到用户。一直以来,电力工业都是世界各国发展经济战略的重中之重,作为一项基础能源产业,它的发展关系到国计民生的发展,是国民经济发展中最重要的一部分[1-2]。然而,近年来,电力行业安全事故依旧时有发生,即使短短几秒钟的一次大停电事故,它所带来的影响也不亚于一场大地震带来的破坏性[3-4]。因此,采取有效的手段来获取并管理电力安全事故信息,仔细分析所有的电力事故发生的原因,充

收稿日期:2018-07-15 **修回日期**:2018-11-15 **网络出版时间**:2019-03-06
基金项目:中央高校基本科研业务费专项资金(2018ZD06)
作者简介:党 佩(1994-),女,硕士研究生,研究方向为计算机应用技术;阎光伟,博士,副教授,研究方向为知识工程、计算机图形图像。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190306.0907.038.html

分发掘事故信息数据的内在价值,为安全事故的预防工作提供参考信息,改善工作中存在的漏洞,才能避免大型事故的再次发生。

互联网的快速发展,使得人们可以应用互联网便捷、高效地获取所需的各种信息。但是,互联网上的网络数据量以惊人的速度呈几何级数增长[5],利用人工获取费时费力,而网络爬虫技术[6-7]是一种快速获取电力事故信息的全新方式。通过利用 WebMagic 增量爬取技术,实现了自动爬取电力安全管理网站中的电力事故信息,并将爬取所获得的结果存储到数据库中。

1 关键技术介绍

1.1 WebMagic

WebMagic 是应用 Java 语言实现的 Web 爬虫,参考了 Scrapy 的设计原理。Scrapy 是一种使用 Python 语言开发,用于快速、高层次的屏幕抓取和 Web 站点抓取,同时能够从抓取的页面中提取出结构化数据的框架[8]。通过使用最成熟的一些 Java 开发工具,如 HttpClient、Jsoup 等完成具体的实现。

WebMagic 是由 Downloader、PageProcessor、Scheduler、Pipeline 四个组件构成的[9],其中 Downloader 完成从互联网上下载页面的功能;PageProcessor 主要是对所要下载的页面进行解析,从中抽取出有用的信息,同时发现一些相关的新的链接;Scheduler 负责管理所有待抓取的 URL,并完成去重的功能;而处理抽取的结果,包括对结果进行计算、将结果持久化到文件或数据库等操作由 Pipeline 组件完成。WebMagic 逻辑的核心组件是 Spider,Spider 可以看作一个大容器,它将其余四个组件管理起来,实现组件之间的相互交互,以及流程化的执行。

图 1 为 WebMagic 总体架构。

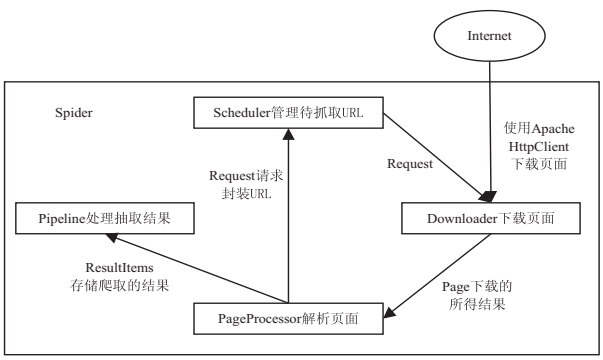

图 1 WebMagic 总体架构

如图 1 所示,Request 既实现了 PageProcessor 与 Downloader 之间的交互,又实现了 PageProcessor 对 Downloader 的控制。Request 通过对 URL 地址进行封装,保证一个 Request 请求对应一个 URL 地址。

通过 Downloader 组件下载得到一个页面,该页面可能是不同格式的文件,如包括 HTML 格式、JSON 格式或其他文本格式等,Page 则用来表示内容,提供抽取信息和保存结果等功能,因此,Page 是实现 WebMagic 抽取功能的核心对象。

PageProcessor 将处理得到的结果交给 ResultItems 对象保存,Pipeline 组件需要使用这些结果时,由 ResultItems 对象提交给它使用。

1.2 Maven

Maven 是一种专业的工具,用于构建和管理 Java 相关项目[10],应用 Maven 管理的 Java 项目都有着相同的项目结构,即:采用统一的标准管理 java 的目录结构,src/main/java 目录下存放所有的 java 代码,而 src/test/java 目录下存放所有的测试代码。同时,Maven 统一维护所有的 jar 包[11],所有需要的 jar 包都被放在一个本地“仓库”里,当某个项目需要用到某一个 jar 包的时候,只需给出当前所需 jar 包的名称和版本号,以此实现 jar 包的共享;当本地“仓库”里找不到所需 jar 包时,根据所提供的名称和版本号,Maven 会自动从远程仓库中下载这个 jar 包,所有的配置只需在 pom. xml 的配置文件中完成。

2 案例实现

文中采用 WebMagic 爬虫技术,以电力安全管理网(http://www. safehoo. com/NewsSpecial/Electric/)为例,获取当前网站中全部与电力事故相关的新闻,包括当前页面的 URL 地址,新闻的标题,以及新闻的内容,并将爬取所得的结果存储进数据库。实验具体的操作流程如图 2 所示。

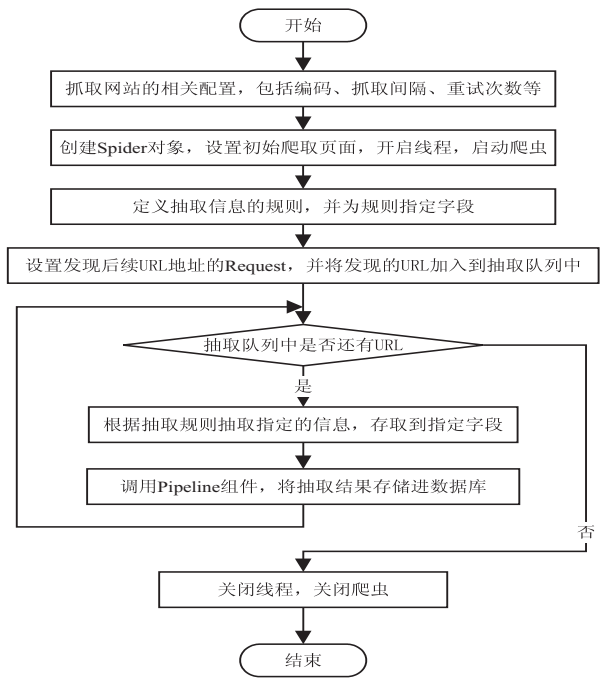

图 2 信息爬取操作流程

2.1　WebMagic 爬虫准备工作

为了应用 WebMagic 进行爬虫操作,首先需要借助 Maven 进行依赖管理,因此在项目的 pom. xml 配置文件中添加对象的依赖即可,如下所示:

```
<dependency>
<groupId>us. codecraft</groupId>
<artifactId>webmagic-core</artifactId>
<version>0. 7. 3</version>
</dependency>
<dependency>
<groupId>us. codecraft</groupId>
<artifactId>webmagic-extension</artifactId>
<version>0. 7. 3</version>
</dependency>
```

2.2　网页信息爬取流程

2.2.1　列表页面及文章页面的 URL 确定

电力安全管理网站是以列表页分页的形式显示电力事故信息的,因此,需要通过遍历这些分页找到所有的目标页面。

通过分析页面 URL,列表页的格式是“http://www. safehoo. com/NewsSpecial/Electric/List _ 1. shtml”,其中 List_1 中的“1”是可变的页数。为了动态获取当前网站总的分页数,进行如下操作:首先,经过分析,当前页面在最下面的分页导航条中,除了提供点击进行跳转到首页,当前页的上一页,当前页的下一页和尾页的功能以外,同时显示当前网站总的数据量和每页可以显示多少条数据的信息的功能。因此,通过使用 HttpURLConnection 类,将当前的页面下载下来,下载的结果即为当前页面的 HTML 编码,将该结果存储到字节流当中,利用正则表达式,匹配出分页导航条中显示的当前网站总的数据量和每页显示多少条数据的信息,并将匹配结果从字节流当中取出,利用这两个数据进行如下判断:

假设 m 表示当前网站总的数据量, n 表示每页可以显示数据的条数, p 表示当前网站的总页数,则 $m \% n \neq 0$,则 $p = m/n + 1$;反之, $m\% n = 0$,$p = m/n$ 。其中,% 表示求余运算,/表示除运算。

以此,可以获取到当前网站的总页数,以总页数为循环条件,依次放入“http://www. safehoo. com/NewsSpecial/Electric/List_ * . shtml”这个 URL 地址中 List_ * 中的“ * ”的位置中,即为列表页的 URL。

而文章页的格式为 http://www. safehoo. com/News/News/China/201103/173316. shtml,其中“201103/173316”是可变的字符串。

2.2.2　定义实体类,封装爬取内容,并解析 HTML

当确定了列表页和文章页的 URL 地址,接下来要先实现对爬取的内容进行确定,解析当前网页的 HTML,获取所需的信息。图 3 所示为打开浏览器的开发者工具后,当前网站的 HTML 结构。

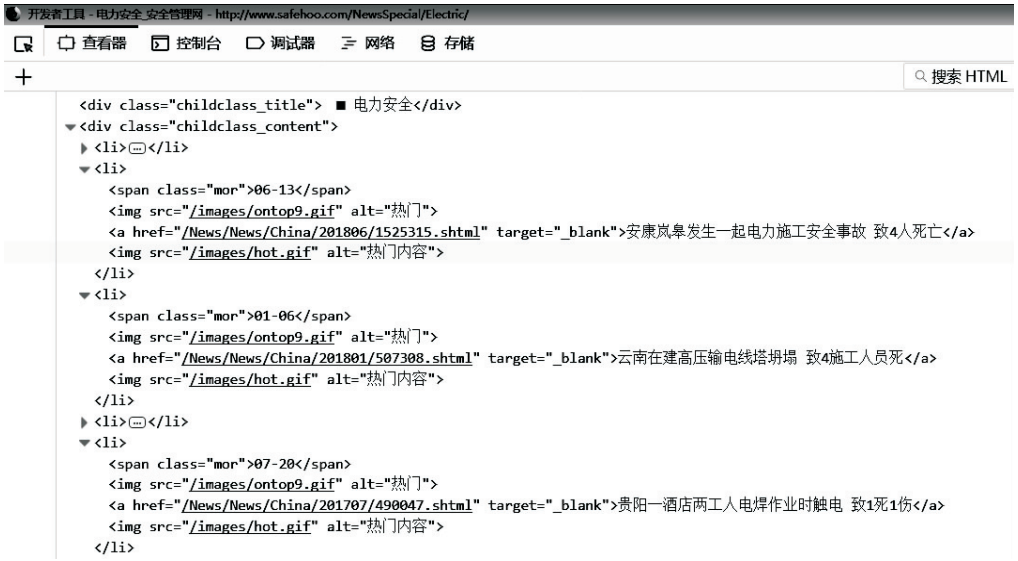

图 3　电力安全管理网 HTML 结构

其中新闻标题存在于 div 的 class 属性为 c_title_text 标签 h1 下,通过 XPath 选择语句//div[@class=´c_title_text´]/h1/text(),即可获得当前的新闻标题。同理,可以分析出新闻内容的选择语句为//div[@class=´c_content_text´]/p/text()。新闻的 URL 利用正则表达式[12]获取,选择语句为 http://www. safehoo. com/News/News/China/\\d+/\\w * \\. shtml。

WebMagic 可以使用注解的方式将抽取规则作用到某个指定的字段上,以此来表示依据这个抽取规则抽取的结果均保存到这个字段中。WebMagic 中提供了两个注解,分别是@ ExtractBy 注解和@ ExtractByUrl 注解。可以使用 CSS 选择器、正则表达式、XPath 和 JsonPath 等方式定义注解表示规则,需要注意,@ ExtractByUrl 注解除正则表达式定义的规则以外,其他

方式均不支持。

利用这两个注解,当前所要爬取的信息实体类定义如图 4 所示。

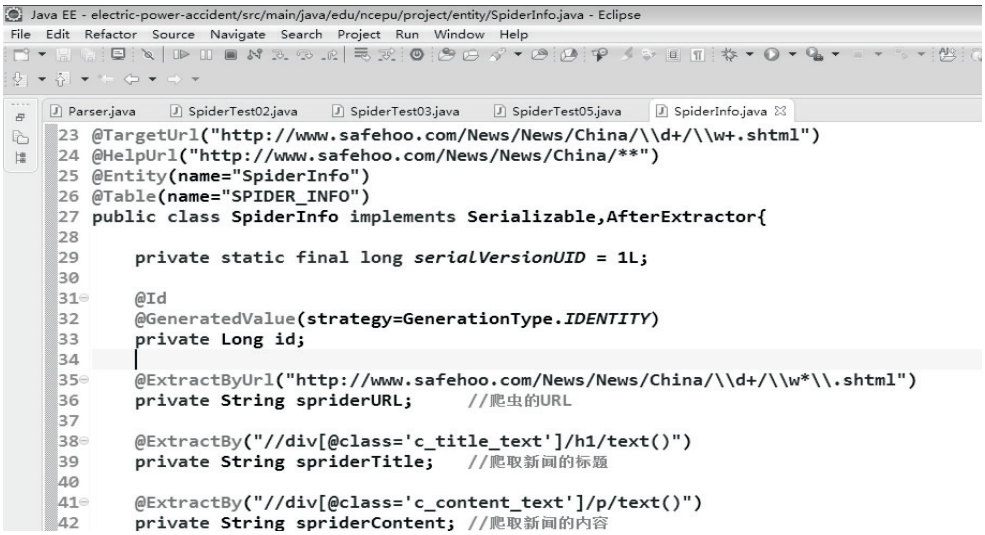

图 4 爬取信息的实体类定义

其中@ TargetUrl 注解表示最终要抓取的 URL,也就是 2.2.1 节分析所得文章页的 URL,最终想要的数据内容都来自与注解中第一的规则匹配的信息,即:http://www. safehoo. com/News/News/China/\\d+/\\w+. shtml;而@ HelpUrl 则是为了帮助找到最终 URL 过程中需要访问的页面,也就是 2.2.1 节分析所得的列表页 URL,即为匹配这个规则(http://www. safehoo. com/News/News/China/**)的所有页面。

2.2.3 定制 PageProcessor

爬虫信息的配置、页面元素抽取规则的制定以及新链接的发现是定制 PageProcessor 过程中的三个主要部分。其中包括对编码信息、每次抓取间隔的时间、超时时间、重试次数等信息的配置,以及一些模拟的参数,比如 User Agent、cookie,以及代理的设置,是爬虫信息配置阶段需要完成的功能。本次实验中,设置重试次数为 2,超时时间为 1 000 s,开启了 2 个线程。表 1 为爬虫配置信息相关的方法。

表 1 爬虫配置信息相关的方法

方法	说明
setCharset(String)	设置编码
setTimeOut(int)	设置超时时间
setRetryTimes(int)	设置重试次数
setUserAgent(String)	设置 UserAgent
addCookie(String,String)	添加一条 cookie

页面元素的抽取,WebMagic 中主要提供了 XPath、正则表达式和 CSS 选择器三种抽取方式。实验中,关于新闻标题和新闻内容的抽取,选择了 XPath 的抽取方式[13];而当前新闻页面的 URL 的抽取,选择了正则表达式的方式,之所以选择不同的抽取方式,因为不同的信息抽取基于不同的注解。

新链接的发现,是指当前所要爬取的一个站点的页面有很多,不可能在开始的时候将全部页面列举出来,因此需要爬虫实现自动发现后续相关链接的功能,如下代码,为本实验中后续代码发现的部分操作。

page. addTargetRequests(page. getHtml(). links(). regex("http://www. safehoo. com/News/News/China/\\d+/\\w*\\. shtml"). all());

其中,page. getHtml(). links(). regex("http://www. safhoo. com/News/News/China/\\d+/\\w*\\. shtml"). all()用于获取所有满足“(http://www. safehoo. com/News/News/China/\\d+/\\w*\\. shtml)”这个正则表达式的链接,而将所有满足条件的链接加入到待抓取的队列中去,以此实现发现后续新的链接的功能由 page. addTargetRequests()方法完成。

2.2.4 编写 Pipeline,将爬取到的数据保存到数据库

Pipeline,是 WebMagic 用于保存抽取结果的组件,其实质是将 PageProcessor 组件抽取得到的结果继续进行后续处理。而之所以采用了两个组件进行数据处理,是基于以下两个原因:第一,实现模块分离,分别将爬虫的两个阶段“页面信息的抽取”和“数据的持久化”交由两个不同的组件完成,既可以保证代码结构清楚,又可以实现处理过程分开进行,完成在独立的线程甚至于不同的机器上执行处理过程的功能;第二,由于 Pipeline 组件实现功能相对固定,保存结果到控制台或者数据库中,这些操作对于所有的页面都是一样的,因此更容易做成通用组件,而页面的抽取方式变化很多,页面的结构也不尽相同,因此抽取规则需要根据每个网页特别定制,不易做成通用组件。

本实验实现了将结果输出到控制台并保存到数据

库中。应用 MySQL 数据库保存抽取的结果,数据持久化层框架[14]采用 JPA(Java persistence API),具体实现过程为:首先,业务逻辑层需要实现 PageModel Pipeline 接口,然后将接口中的抽象方法 process(ResultItems resultItems,Task task)重写,其中参数 ResultItems 类用来保存抽取结果,它本身是键值对的结构,可以通过 ResultItems. get(key)获取在 page. putField(key,value)中保存的数据。该方法中添加输出到控制台的语句,并且调用数据持久化层保存数据进数据库的方法即可。图 5 所示为保存到数据库中的结果。

对象　spider_info @electric (accid...　spider_info @electric (accid...

开始事务　备注　筛选　排序　导入　导出

id	sprider_content	sprider_title	sprider_url
1	记者从云南省麻栗坡县委宣传部获悉，1月5日16时许，麻栗坡	云南在建高压输电线塔坍塌 致4施工人员	http://www.safehoo.com/News/News/China/20
2	据国家能源局网站消息，国家能源局近日通报2017年10月全国	10月全国发生电力人身伤亡责任事故5起	http://www.safehoo.com/News/News/China/20
3	前几天，在陕西蒲城一家西北地区最大的电厂里发生了一起事	陕西渭南蒲城一电厂隐瞒安全事故 两人不	http://www.safehoo.com/News/News/China/20
4	6月19日，澎湃新闻（www.thepaper.cn）从湖南永州市委宣	湖南永州一变压器爆炸致路人1死7伤	http://www.safehoo.com/News/News/China/20
5	3月21日上午，网友爆料徐州丰县发生一起意外事故，两名电	江苏丰县发生一起供电作业工人触电事故	http://www.safehoo.com/News/News/China/20
6	11月8日上午9:43许，山东淄博市周村嘉周热电有限公司脱硫	山东淄博热电厂发生爆炸 已致5死6伤	http://www.safehoo.com/News/News/China/20
7	据国家安监总局消息，8月11日15时20分许，湖北省宜昌市当	湖北当阳一电厂发生爆炸事故 致21死5伤	http://www.safehoo.com/News/News/China/20
9	7月23日11时30分左右，北京森桦建业防水工程有限公司工人	河北故城工人施工触电 3死1伤	http://www.safehoo.com/News/News/China/20
10	7月5日下午，永州市冷水滩区高溪市镇青山洞村横冲组，发生	湖南冷水滩发生致3人死亡农田电击事故	http://www.safehoo.com/News/News/China/20
11	据广东省应急办通报,2015年10月21日上午10:30，茂名市滨海	广东茂名滨海新区发生一起触电事故造成	http://www.safehoo.com/News/News/China/20
12	6月18日，上栗县金山镇昌栗高速公路C12标段发生一起意外触	江西上栗县发生一起意外触电事故 造成2	http://www.safehoo.com/News/News/China/20
14	7日上午，潍坊市民致电大众网，称高密市物流园6日发生触电	潍坊高密姜庄物流园发生触电事故致4人伤	http://www.safehoo.com/News/News/China/20
15	记者5月4日从安徽省安监局获悉，5月3日15:45许，舒城县阚	安徽舒城发生输电工程事故致3人死亡	http://www.safehoo.com/News/News/China/20
16	4月4日下午，在张店南部一厂房院内发生一起触电事故。几名	淄博张店一厂房发生触电事故 3人身亡	http://www.safehoo.com/News/News/China/20
17	11月15日讯 据国家能源局网站消息，2014年10月，全国发生	10月全国发生电力人身伤亡责任事故3起	http://www.safehoo.com/News/News/China/20
18	10月30日上午11时左右，合肥市包河区派河路与宿松路口西北	合肥一工地违规操作酿惨祸 施工人员触电	http://www.safehoo.com/News/News/China/20
19	12日上午9时30分左右，江西省萍乡市湘东区培兴精制酱菜厂	江西萍乡酱菜厂发生漏电事故致3死4伤	http://www.safehoo.com/News/News/China/20

图 5　数据库存储结果

3　结束语

通过对 WebMagic 爬虫技术的研究,应用 Maven 技术管理 java 项目,实现了对电力安全管理网站上电力事故信息的爬取和存储。在实验过程中,通过对网页结构的分析,分别对所要爬取的字段采用 XPath 和正则表达式的方式指定抽取规则,定制 PageProcessor 组件,实现信息的顺利爬取,同时,应用 Pipeline 组件,实现爬取数据存储入库。应用该模式,可以实现对其他相关电力网站事故信息的爬取,通过整理分析历年事故信息,对每次事故发生原因高度重视,采取必要的措施加以解决和管理,避免大型事故的重复发生。只有遏制住电力生产安全事故的发生,才能创造出更好的经济效益。

参考文献:

[1] 刘保国,林　方. 贯彻科学发展观与促进电力产业发展[J]. 工会论坛:山东省工会管理干部学院学报,2012(6):88-90.

[2] 何国才. 新形势下电力安全生产管理思考与探索[J]. 企业技术开发,2017,36(2):137-139.

[3] 李会俊. 关于电力安全监察工作问题探析[J]. 民营科技,2014(11):117.

[4] BATRA I. Electric accidents in the production,transmission,and distribution of electric energy:a review of the literature[J]. International Journal of Occupational Safety and Ergonomics,2001,7(3):285-307.

[5] 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):199-202.

[6] 于　娟,刘　强. 主题网络爬虫研究综述[J]. 计算机工程与科学,2015,37(2):231-237.

[7] STEVANOVIC D,AN Aijun,VLAJIC N. Feature evaluation for web crawler detection with data mining techniques[J]. Expert Systems with Applications,2012,39(10):8707-8717.

[8] 李代祎,谢丽艳,钱慎一,等. 基于 Scrapy 的分布式爬虫系统的设计与实现[J]. 湖北民族学院学报:自然科学版,2017,35(3):317-322.

[9] 武婷婷. 一种基于 WebMagic 和 Mahout 的信息搜集与推荐系统[J]. 软件导刊,2016,15(10):1-3.

[10] 杨新艳,于伟涛. 基于 Maven 的轻量级 Java 软件开发研究[J]. 科技传播,2015,7(17):134-135.

[11] 董晓光,喻　涛. 使用 Maven 构建 java 项目[J]. 电子技术与软件工程,2014(10):105.

[12] 胡军伟,秦奕青,张　伟. 正则表达式在 Web 信息抽取中的应用[J]. 北京信息科技大学学报,2011,26(6):86-89.

[13] 阮　娟. 基于 XPath 的新闻信息抽取系统设计与实现[J]. 智能计算机与应用,2015,5(2):58-61.

[14] 李华勇. 计算机数据库存储技术的开发与应用[J]. 长沙铁道学院学报:社会科学版,2013,14(2):199-200.