基于 Scrapy 的商品评价获取系统设计*

施　威,夏　斌
(上海海事大学 信息工程学院,上海 201306)

摘　要:随着电子商务的迅速发展和竞争愈加激烈,对于电商平台上第三方卖家而言,如何准确获取商品评论信息从而正确选择上架的商品变得愈来愈重要。目前第三方卖家在获取商品评价工作上主要依赖于人工收集信息,不仅效率十分低下,并且准确度得不到保障。为了帮助电商平台上第三方卖家高效并准确地解决这一问题,文中设计出了一种基于网络爬虫的商品评价获取工具。该工具实现了对一个畅销商品类目下的所有商品评论进一步细化与筛选,为用户提供更加直观的商品指标,同时固化存储商品评论为后续的进一步优化提供数据源。该系统主要技术采用 Scrapy 框架,开发语言采用 Python2.7,经过测试后发现达到了良好的效果。

关键词:电子商务;网络爬虫;Scrapy;Python
中图分类号:TP391.9　　**文献标识码**:A　　DOI:10.19358/j.issn.1674-7720.2017.19.004
引用格式:施威,夏斌. 基于 Scrapy 的商品评价获取系统设计[J]. 微型机与应用,2017,36(19):12-15.

The design of product reviews acquisition system based on the Scrapy framework

Shi Wei, Xia Bin
(College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

Abstract: With the rapid development of e-business and the increasingly fierce competition, how to accurately obtain product reviews information and select stores goods correctly become more and more important, especially for those third-party sellers on e-business platform. At present, the third-party sellers mainly depend on artificial gathering information for the product reviews getting work, not only the efficiency is very low, and the accuracy can not be guaranteed. In order to help those third-party sellers on e-business platform to solve this problem efficiently and accurately, this paper designed a product reviews acquisition tool based on web crawler. This tool implements further refining and filter for all goods under a best-selling merchandise category, to provide users with more intuitive product indicators, at the same time to grab and store goods comments as data source for further optimization. The main technology of this system is Scrapy framework, development language adopts Python 2.7. After testing, it has achieved good results.

Key words: e-business; the crawler; Scrapy; Python

0　引言

电子商务的兴起促进了商业模式的变革,作为最资深的电商平台,Amazon 拥有巨大的用户群体,仅仅美国站的第三方卖家数量就超过 20 万。平台上每个商家店铺的运营好坏与商家的选品质量紧密相关,尤其对于平台上的第三方卖家。选品的效率越高,准确度越高,商品自然更加吸引消费者,店铺的客户群体也会更多。因此提高选品质量是提升店铺收益的重要手段。

目前选品工作的一个重要判断依据就是商品的评价信息,如何高效准确地获取商品评价信息并得出一些商品的相关数据指标,对于选品的质量至关重要。获取商品评价信息主要依赖于人工去检索信息,这种方式效率十分低下。另外电商平台也会提供商品综合评价指标,但数据量太过抽象,可参考性不足。对于不同商家对评论数据的需求不同,电商平台很难提供有价值的信息。因此本文设计了一个基于 Scrapy 框架的评价获取系统,用户通过提供特定的商品类目来获取更加直观的商品评价数据,从而为选品工作提供相应评价指标。

1　相关技术简介

1.1　网络爬虫与 Python

网络爬虫[1](Web Crawler)是一种特定的应用程序或者脚本,可以按照一定的匹配规则自动地提取 Web 页面中特定的内容。它最典型的应用就是搜索引擎从互联网上抓取数据,并且下载 Web 页面。网络爬虫最原始的目的就是从互联网上下载数据到本地进行备份。爬虫是从一个或有多个 URL 的集合开始进行爬取,首先获取一个 URL 并下载此 URL 页面内容,提取该页面中其他需要的

*基金项目:上海市科学技术委员会资助项目(14441900300);国家自然科学基金(61550110252)

URL 放入集合队列中,反复此过程直至爬取所有 Web 页面。常见的爬取策略有广度优先爬虫、重复爬取已有页面爬虫和定向爬虫。

Python 语言是一种语法简单明晰、功能强大、兼具面向过程与面向对象的开源编程语言,特别适用于应用程序的敏捷开发,它几乎可以在所有主流的操作系统上运行。Python 语言提供了非常丰富的网络协议标准库,例如自带的 urllib、urllib2 等最基本的爬虫库。另外,Python 生态包含非常丰富的第三方工具包[2],比如强大的 Scrapy、requests、BeautifulSoup 等网络工具库。

1.2 Scrapy

Scrapy[3]是基于 Python 语言开发的一个开源 Web 并行爬取框架,它能够快速爬取 Web 站点并从页面中提取自定义的结构化数据。因突出的爬取性能,Scrapy 在数据挖掘、数据监测和自动化测试领域得到了广泛应用。Scrapy 使用 Twisted 这个异步网络库来处理网络通信,架构清晰,并且包含了各种中间件接口,用户只需要在 Scrapy 框架的基础上进行模块的定制开发就可以轻松实现一个高效的爬虫应用[4]。Scrapy 整体架构如图 1 所示。

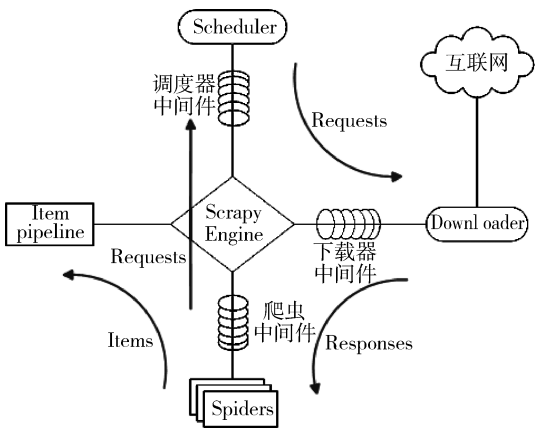


图 1 Scrapy 整体架构

(1)Scrapy Engine:框架引擎,用来处理整个系统的数据流处理,触发事务;

(2)Scheduler:调度器,用来接受 Engine 发过来的请求,压入队列中,并在引擎再次请求时返回;

(3)Spiders:蜘蛛,也称为爬虫,用来定制特定解析规则爬取页面并提取自定义的 item 数据;

(4)Downloader:下载器,用来下载页面内容,并将内容返回给 Spiders;

(5)Item Pipeline:项目管道,Spiders 解析过后的数据被送到项目管道进行进一步的处理;

(6)Downloader Middlewares:下载器中间件,处理 Scrapy 引擎与下载器之间的请求及响应。

1.3 Xpath

Xpath 即为 XML 路径语言,它被用来标示 XML(标准通用标记语言的子集)文档中的特定位置[5]。Xpath 基于 XML 的树状结构,提供在树形结构数据中定位节点的功能。因为爬虫爬取的通常是 HTML 页面,HTML 同 XML 一样也是树状结构,所以 Xpath 同样也支持 HTML。爬虫的目的是为了获取数据,而需要的数据通常都不是页面的全部,获取指定的数据需要进行数据匹配。常用的匹配技术有 Python 自带的正则表达式类库(re),但正则匹配不能完全保证匹配到指定的数据节点,表达式的书写也比较复杂。Xpath 语言简化了匹配表达式的书写,匹配成功率也更高。Python 语言对 Xpath 具有良好的支持,Scrapy 与 Xpath 结合使得爬虫效率高效并且可靠。

2 系统设计

2.1 系统整体框架

系统主要分为五大部分,分别是 URL 管理、页面数据解析、数据的提取与处理、数据的存储、爬虫调度。Scrapy 引擎从 URL 管理器获取需要并行爬取的 URL,然后进行页面解析,提取出需要的数据。再进一步进行数据过滤及处理,最后将需要的数据进行存储固化,将结果导出为 Excel 表格。系统整体框架如图 2 所示。

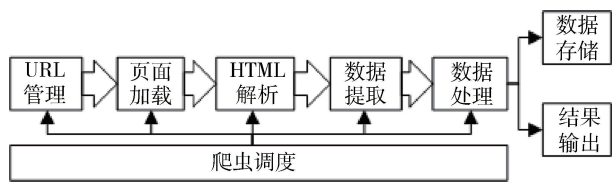


图 2 系统框架

2.2 爬虫调度

爬虫调度[6]是 Scrapy 爬虫的控制单元,对爬虫系统的各个模块进行协调和调度,核心功能包括:

(1)实现抓取数据的流程;

(2)控制其他模块的执行;

(3)为 HTTP 请求添加请求信息,例如 Headers;

(4)采取合适的反爬虫措施避免反爬虫机制。

2.3 页面加载器

页面加载器通过爬虫调度器提供的 HTTP 请求体信息以及 URL 管理器的 URL,向 Web 服务器发起 HTTP 请求,获取服务器响应的 HTML 页面。为了避免反爬虫机制导致服务器无法及时响应或拒绝访问,页面加载器采用定时机制限制请求的频次。

2.4 HTML 解析器

HTML 解析器对页面加载器获取的页面数据进行解析,解析后的数据为树状结构,以便后续利用 Xpath 进行匹配选取。同时 HTML 解析器会将解析出来有需要的 URL 反馈给爬虫调度器。本系统采用的是第三方 Python 网络协议库:requests,它包含的 get、post 等静态方法对常见的 HTTP 请求响应处理都有很好的封装。

2.5 数据输出

数据输出分为两部分,一部分是将有用的数据固化下来,以便后续对系统进一步优化升级。本系统采用 Python 自带的 IO 文件模型进行固化。另一部分是将系统计算的结果导出为 Excel 表格,采用 Scrapy 自带导出命令行工具,简单直观。

3 系统实现

3.1 页面解析及数据提取

页面加载过后需要对页面数据进行解析,解析后提取出需要的数据,然后对提取出的数据进行整合,最后对数据加以处理,并将有效的评价信息进行文件固化。利用 Xpath 对 HTML 原数据进行提取操作。例如,提取商品价格 get_price 的 Xpath 语法如下:link = get_price.xpath('div[@class="zg_itemWrapper"]/a[@class="a-link-normal"]/@href')[0].extract()。同理也能得到其他需要的数据,然后根据定义的数据模型进行整合。Item 是 Scrapy 用来保存数据的容器模型,创建 Item 子类并定义相应字段 field 即可创建数据模型。数据提取流程如图 3 所示。

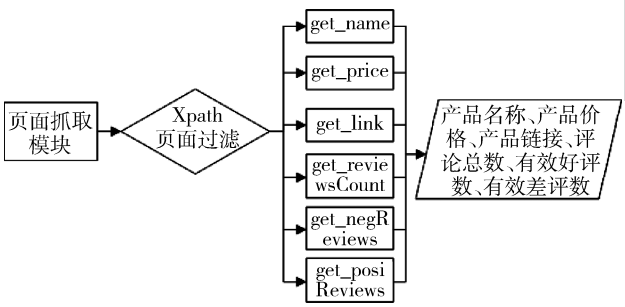


图 3 数据提取逻辑图

3.2 数据的处理及输出

在提取到商品基本信息(名称、价格、链接)后,通过商品详情页的 URL 再次请求得到商品的评价数据。Scrapy 是一个异步的爬虫框架,所以异步对评价进行深度遍历增强了程序的交互友好性。Amazon 的评价分为两种,一种是验证通过(Verified Purchases)的评论,一种是未通过的。未通过 Amazon 认证的评价显然可靠性不高,所以此类评价将被过滤掉。数据处理逻辑如图 4 所示。

评价时效区间的选择可以自定义设定,一般认为近 6 个月的评价更具参考性。评价的好差评定策略以 3 星为分界点,大于 3 星认为是好评,小于等于 3 星认定为差评。评价的固化利用 Python 的 IO 文件模型保存到 txt 文件,文件名为商品名,以便后期对评价的内容做进一步的分析。输出结果利用 Scrapy 的结果输出模块导出为 Excel 表格,如图 5 所示。

proReviewSum 为所有认证过的评价总数,proPosiReview 为有效时间段内好评总数,proNegReview 为有效时间段内

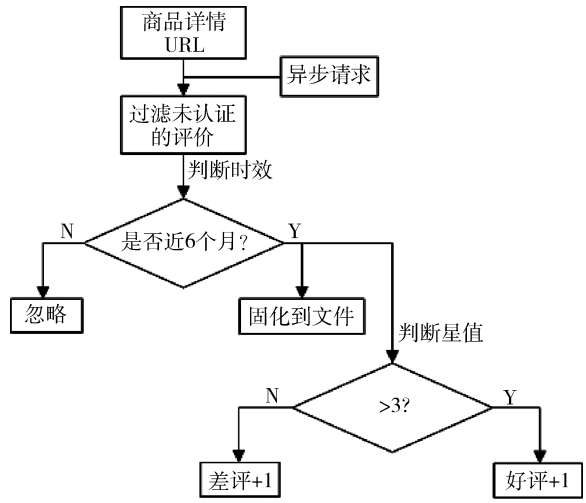
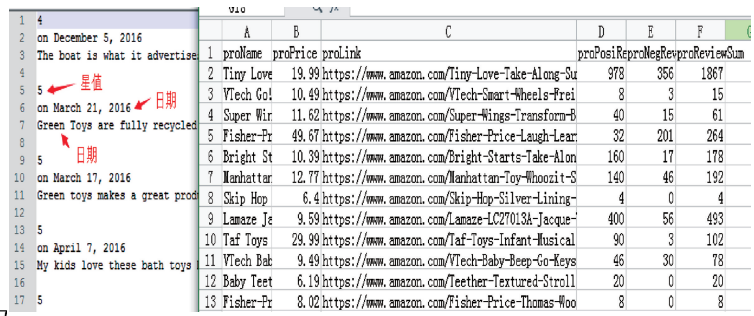


图 4 数据处理逻辑图



	A	B	C	D	E	F
1	proName	proPrice	proLink	proPosiRe	proNegRe	proReviewSum
2	Tiny Love	19.99	https://www.amazon.com/Tiny-Love-Take-Along-Su	978	356	1867
3	VTech Go!	10.49	https://www.amazon.com/VTech-Smart-Wheels-Frei	8	3	15
4	Super Wir	11.62	https://www.amazon.com/Super-Wings-Transform-B	40	15	61
5	Fisher-Pr	49.67	https://www.amazon.com/Fisher-Price-Laugh-Lear	32	201	264
6	Bright St	10.39	https://www.amazon.com/Bright-Starts-Take-Alon	160	17	178
7	Manhattar	12.77	https://www.amazon.com/Manhattan-Toy-Whoozit-S	140	46	192
8	Skip Hop	6.4	https://www.amazon.com/Skip-Hop-Silver-Lining-	4	0	4
9	Lamaze Je	9.59	https://www.amazon.com/Lamaze-LC27013A-Jacque-	400	56	493
10	Taf Toys	29.99	https://www.amazon.com/Taf-Toys-Infant-Musical	90	3	102
11	VTech Bak	9.49	https://www.amazon.com/VTech-Baby-Beep-Go-Keys	46	30	78
12	Baby Teet	6.19	https://www.amazon.com/Teether-Textured-Stroll	20	0	20
13	Fisher-Pr	8.02	https://www.amazon.com/Fisher-Price-Thomas-Woo	8	0	8

图 5 样例展示

差评总数。根据结果数据可以得出两个参数,一个为总的评价数量,数量越高说明该产品人气越高,另一个为好评数与差评数的比值,比值越高说明产品好评率越高。对两个参数进行综合考量,可以对选品起到一定的指导参考作用。

3.3 反爬机制应对策略

反爬[7]是很多网站都会采取的保护措施,不同网站采取的反爬策略不同,所以本系统实现了常见的反爬应对策略,充分保证系统运行不受反爬机制的影响。实现的反爬应对策略如下:

(1)设置 DOWNLOAD_DELAY,该参数为 Scrapy 在同一个网站两个不同页面之间跳转需要等待的时间。可以在 setting.py 文件里面设置:DOWNLOAD_DELAY = 3 s;

(2)使用用户代理 User-Agent,User-Agent 是描述 HTTP 请求终端信息的参数,使用动态变化的 User-Agent 可以避免反爬机制的识别和访问流量统计异常。采用 User-Agent 池加上 Python 原生随机生成算法可以实现动态变化的 User-Agent;

(3)禁止 cookies[8],可以防止网站利用 cookies 识别爬虫轨迹[9-10]。在 setting.py 文件中设置:COOKIES_ENABLED = False。

4 结论

通过将商品基本信息、评论数据抓取下来,并对评论进一步地精确化,计算出更具参考价值的选品指标,对选

品质量起到了一定的提高作用,同时也大大节约了众多店铺商的手工查询时间,帮助他们实现更好的收益。本文利用互联网技术简化了电子商务平台上繁杂性的工作,有很强的应用价值。

参考文献

[1] PANI S K, MOHAPATRA D, RATHA B K. Integration of Web mining and Web crawler: relevance and state of art[J]. International Journal on Computer Science & Engineering, 2010, 2(3):772-776.

[2] 徐咏梅. Python 网络编程中的远程调用研究[J]. 电脑编程技巧与维护, 2011(18):80-81.

[3] XIE D X, XIA W F. Design and implementation of the topic-focused crawler based on scrapy[J]. Advanced Materials Research, 2013, 850-851:487-490.

[4] Twisted 15. 4. 0 documentation [EB/OL]. (2017-02-20) [2017-03-26]. http://twistedmatrix. com/documents/current/core/howto/defer. html.

[5] 杨文柱, 徐林昊, 陈少飞,等. 基于 XPath 的 Web 信息抽取的设计与实现[J]. 计算机工程, 2003, 29(16):82-83.

[6] 李婷. 分布式爬虫任务调度与 AJAX 页面抓取研究[D]. 成都:电子科技大学,2015.

[7] 邹科文,李达,邓婷敏,等. 网络爬虫针对"反爬"网站的爬取策略研究[J]. 电脑知识与技术,2016(7):61-63.

[8] HARDING W T. Cookies and Web bugs: what they are and how they work together[J]. Information Systems Management, 2001, 18(3):17-24.

[9] 漆志辉, 杨天奇. 网络爬虫性能研究[J]. 微型机与应用, 2011, 30(5):72-74.

[10] 王勇杰. 电子商务网站中购物车的实现[J]. 微型机与应用, 2011, 30(17):11-12.

(收稿日期:2017-04-08)

作者简介:

施威(1993 -),男,硕士研究生,主要研究方向:智能商务信息处理。

夏斌(1975 -),通信作者,男,博士,副教授,硕士生导师,主要研究方向:脑-机接口、云计算及人工智能。E-mail:xawen267@gmail. com。

(上接第 8 页)

6 结论

本文从提高对变化车流量调度效率的角度出发,设计了基于 CAN 总线技术的交通管理控制系统。经测试,该系统能够根据车流量信息智能调节车辆通行时间,提高了车辆通行效率。系统模块之间采用 CAN 总线通信,通信速率高、稳定性好。整个系统采用模块化设计思想,方便了系统的维护。能够根据排队车辆数预测车辆的通行时间。本系统安全可靠,参数配置方便,具有一定的推广价值。

参考文献

[1] 郭继孚, 刘莹, 余柳. 对中国大城市交通拥堵问题的认识[J]. 城市交通,2011,9(2):8-14,6.

[2] 金茂菁. 我国智能交通系统技术发展现状及展望[J]. 信息与安全,2012(5):1-5.

[3] 杨飞,郑贵林. 基于 CAN 总线的监控系统设计[J]. 微计算机信息,2005,21(7):34-36.

[4] 喻金钱,喻斌. STM32F 系列 ARM Cortex-M3 核微控制器开发与应用[M]. 北京:清华大学出版社,2011.

[5] 孙书鹰,陈志佳,寇超. 新一代嵌入式微处理器 STM32F103 开发与应用[J]. 微计算机应用,2010,31(12):59-63.

[6] 陈健. 基于 AWA14423 探头的数字声级计的研制[D]. 哈尔滨:哈尔滨工业大学,2011.

[7] 曹晓伟. MOC3061 系列光电双向可控硅驱动器[J]. 国外电子元器件,1996(12):2-4.

[8] 王毅峰, 温希东. 基于 CAN 总线的智能控制器的设计[J]. 仪表技术与传感器, 2006(4):32-34.

[9] 颜自勇. CAN 总线技术在智能楼宇通信中的应用[J]. 安防科技,2006(4):18-19,54.

[10] 熊茂华. 基于 CAN 现场总线的智能交通信号控制系统[J]. 工业控制计算机,2006,19(9):56-57,64.

[11] 黄波土. 基于 CAN 总线的集散控制系统的研究与设计[D]. 济南:山东科技大学,2003.

[12] 冯亚军. 重型汽车 CAN 总线控制系统的应用研究[D]. 济南:山东大学,2006.

(收稿日期:2017-05-02)

作者简介:

王勇(1970 -),男,博士,工程师,主要研究方向:电子技术。

郭美春(1976 -),女,本科,主要研究方向:光电技术。

佟国栋(1990 -),通信作者,男,硕士研究生,主要研究方向:嵌入式系统。E-mail:745093889@qq. com。

(上接第 11 页)

[13] CHINI P, GIAMBENE G. QoS in hybrid WiFi and DVB-RCS networks[C]. Wireless Communication Systems. 2008. ISWCS' 08. IEEE International Symposium on. IEEE, 2008:718-722.

[14] ALENA R, NAKAMURA Y, FABER N, et al. Heterogeneous spacecraft networks: wireless network technology assessment[C]. IEEE Aerospace Conference, 2014:1-13.

(收稿日期:2017-04-05)

作者简介:

胡伟(1990 -),男,硕士研究生,主要研究方向:网络 MAC 层协议。

陶孝锋(1978 -),男,硕士,研究员,主要研究方向:卫星通信。