文章编号:1007-1423(2018)11-0064-04 DOI:10.3969/j.issn.1007-1423.2018.11.013

基于自动化测试的反爬虫技术研究——以天猫平台为例

曹文斌,张科静

(东华大学旭日工商管理学院,长宁 200051)

摘要:

随着大数据、云计算、移动互联网等新兴信息技术地快速兴起,人们工作生活对互联网的依赖逐步加强,越来越多用户行为数据、发表意见评论数据分散在互联网上。如何有效地采集这些数据,是分析、挖掘数据的前提。传统爬虫方式通常是从一个节点开始,盲目地、发散式地扩张遍历网页来获取数据,这种方式一方面近年来受到反爬虫技术的限制,另一方面获取数据的效率和质量偏低。在深入研究天猫平台网站结构的基础上,采用自动化测试技术模拟人浏览网页的方式,结合电商平台的搜索引擎有效地避开反爬虫技术地限制,采集到的数据准确率达到96%以上,能够满足实际科研、工业中数据采集分析的要求。

关键词:

反爬虫; 自动化测试; 评论; Selenium

0 引言

伴随着互联网信息技术的发展,特别是UGC(User Generate Content)技术的兴起,互联网上信息呈现爆发式增长,各类信息杂乱交错地分散在各大平台。有效地收集整理这些多维、非结构化数据,是各大科研机构、互联公司实现挖掘算法、挖掘"数据资产"的首要前提。

通过传统的爬虫机制,自定义数据获取规则,能够在一定程度上满足个性化需求、获得主题相关的数据,这在过去能够满足大部分的科研和生成的需求。然而,近年来社会步入大数据时代,越来越多的大平台公司意识到数据是一家公司的重要资产。越来越多的公司投入大量的人力和物力构建反爬虫屏障,通过反爬虫技术区分正常使用的用户访问请求和通过爬虫软件密集地、反复地、大批量地访问请求。因此,通过传统的爬虫手段获取数据越来越困难,如何绕开这些反爬虫的限制是快速获取批量数据的一个难题。

本研究结合自动化测试技术和天猫平台的搜索引擎,不仅克服了天猫平台反爬虫机制带来的各种障碍,而且获取的数据相关度较传统的爬虫手段获取的数据要高。同时通过分析JSON数据结构,获取的数据维度超过传统的搜索引擎获取的数据维度。数据的"多维性"对数据的进一步挖掘分析具有重要的作用。

1 爬虫软件的通常思路

网络爬虫(Web Crawler)是一种能够根据事先设定的规则,自动地在网络上像蜘蛛网络一样扩展并采集数据的运用程序或网页脚本。人们也形象称之为网络蜘蛛,该程序能够向蜘蛛织网一样不断扩展网页链接,并遍历采集网页内容,是搜索引擎中必不可少的组成部分。爬虫程序信息采集过程一般是从一个初始超链接(URL)集合,又称种子集合开始,首先把种子集合放入待抓取数据的任务队列中,再根据事先设定好的规则从中取出并访问超链接,遍历超链接指向的网页分析提取数据,同时若发现新的链接则提取该链接放入待抓取数据的任务队列中。再按规则取出任务队列中的下一个链接,如此循环往复直至任务队列为空或者满足设定的终止条件,实现遍历"蜘蛛网"的效果[1-2]。网络爬虫可以大致分为四种,分别是:聚焦网络爬虫、深层网络爬虫、增量式网络爬虫、通用网络爬虫四种,本课题的研究是一种增量式网络爬虫。增量式网络爬

虫(Incremental Web Crawler)仅对已遍历过的网页中最新变动(产生新的内容或已有内容的更新)的部分进行数据分析提取,达到增量式效果[3],相对于通用型爬虫程序具有更强的针对性,降低了访问的盲目性。本课题每次爬取时仅抓取新增最近的新评论,避免抓取以前下载过的评论。

随着信息技术的蓬勃发展,出现了各类爬虫工具、爬虫算法,各类爬虫技术都有自身的特点,"以假乱真"的策略也很多,如模拟登录、动态 IP 等,但爬虫程序抓取数据的过程基本思路大体如图 1 所示:

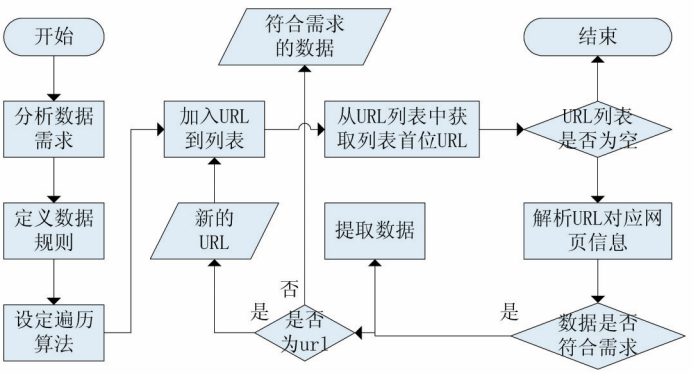

图1 爬虫机制的通常思路

(1)首先要调研了解数据需求。爬虫是一个自动扩展遍历网页的过程,而网页上的信息通常是多种多样的、非结构化的,因此先要了解需求即明确爬虫方向;

(2)根据步骤(1)中的需求定义数据规则。在本课题中是试用正则表达式来定义需求数据的规则,正则表达式是由一些特定的字符组成的规则字符串,规则字符串可以对网页文本进行查找匹配、过滤筛选得到想要的数据;

(3)设定遍历搜索的规则。爬虫的本质是一次查找遍历过程,遍历的算法通常有深度优先、广度优先等算法。我们可以把所有的网页布局当作一幅有向图,每个网页相对于一个节点,超链接相对于有向图的边。这样就能根据图的深度优先或者图的广度优先进行遍历搜索,在本课题中采用的是广度优先搜素规则;

(4)根据数据规则匹配网页文本。这是整个过程中最为核心的步骤,根据步骤(2)中定义的数据规则去查找网页中的内容,匹配到的数据若为需要的数据,则提取存储到本地数据库里面。若为新的超链接,则把新发现的超链接加入到 URL 列表中,这样就能像蜘蛛一样展开。

2 天猫反爬虫机制阻扰

反爬虫是使用相应的信息技术措施,阻止爬虫程序批量获取自己网站信息的一种方式。爬虫和反爬虫机制是一个对立统一的关系,相互依存、相互作用、相互转换。反爬虫机制因爬虫技术而起,反爬虫机制是应对爬虫技术,进行数据保护的措施。在实践过程中,爬虫和反爬虫技术相互促进,爬虫技术的改进必然会驱使人类探究新的反爬虫技术;同样,新的反爬虫技术出现必然促使数据保护方研究新的反爬虫技术。

经实验观察发现,阿里巴巴天猫商城平台基于安全考虑,相对于京东商城而言,平台数据保护措施更为严格。在目前阶段,各大电子商务平台甚至包括国内在线旅行社 OTA(Online Travel Agent)平台、O2O 平台中,天猫商城平台数据保护措施是最为严格的。主要有以下几点:

(1)能够通过鼠标操作路径等方式区分真实用户和爬虫程序的访问;

(2)对不同重要性的信息设置了不同级别的安全应对机制;

(3)能够有选择性地开放给搜索引擎(如百度、谷歌)检索,百度公司都不能全面抓取天猫到的商品信息;

(4)天猫平台对于关键信息,如评论、品属性信息和搜索返回信息保护极为严格,都是动态生成数据,返回浏览器经过浏览器解析才行。返回的结果不能直接显示在网页上,需要在浏览器本地执行 JavaScript 发送请求到服务器,再返回 JavaScript 再浏览器本地执行,多次轮回执行结果拼装组合而成。

3 自动化测试 Selenium

自动化测试,就是再总结测试人员日常操作之后设计开发程序或者在测试工具中设定规则,启动程序或工具能够模拟人的操作,进而控制测试过程中的各种对象和类,达到辅助测试的效果,减轻测试人员的重复性工作[4]。软件自动化测试能大大提高生产效率,提高测试的覆盖率及可靠性,是手工测试的一种有益的

补充。

现有的自动化测试框架不少，不同的框架有不同的特点。本课题根据课题需要选择了 Selenium 框架，相对于其他测试框架而言，Selenium 有自己的优势：

（1）仿真性强，能够像真实用户操作一样，直接在浏览器中运行；

（2）使用方便，提供 API 接口，供 Java 等多种高级程序设计语言调用；

（3）Selenium 核心 browser bot 是用 JavaScript 实现的，有助于避开反爬虫技术的限制。

本课题在充分了解天猫反爬虫机制之后，结合浏览器自动化测试框架 Selenium，编写的 Java 评论采集程序，达到了模拟人的操作浏览器的效果。能够通过 Java 程序自如地启动关闭浏览器、切换浏览的网页、输入数据到网页控件中、点击网页上的按钮等操作，通过浏览器自如地和天猫平台进行数据交互。

4 评论数据并发的获取

本课题编写的 Java 评论采集程序，突破了反爬虫技术限制，解决了批量评论采集的难题。程序的运行逻辑如图 2 所示：

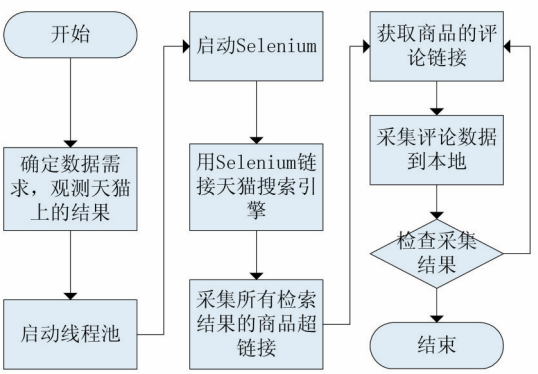

图2 基于Selenium并发分布式抓取数据

第 1 步：首先在天猫上搜索目标商品，并记录搜索结果页码数量和显示搜索结果的 URL。搜索结果的 URL 含有搜索关键词信息和搜索结果显示页面的页码序号，根据页码递增能够遍历所有的搜索结果；

第 2 步：把第一步记录的 URL 和总页码数设置为程序参数，并根据需要采集数据量的大小设置线程池的最大线程数量，启动程序运行；

第 3 步：保存搜索引擎搜索的结果。程序会启动 Selenium 像人工操作一样快速遍历所有搜索结果，并记录每个搜索结果商品的 URL、单价、评价数、月销售量，保持至本地数据库；

第 4 步：根据搜索结果采集评论。根据第 3 步采集的商品 URL 采集该商品的评论，保持至本地数据库。返回的评论数据是 JSON 类型，里面包含的商品评论属性主要包括三大部分，分别是评论属性集、顾客属性集、交易属性集。

5 评论数据的采集结果

本课题研究的实例是阿里巴巴天猫平台，设计的程序在天猫上采集的每条网络购物评论数据包含 38 个属性，这些属性可分为 3 个部分，评论属性集、顾客属性集、交易属性集。每个属性集的主要信息如下面的列表：

表 1 基于 Selenium 的网络评论采集结果之评论属性集

商品评论	rateContent	服务评论	serviceRateContent
追加评论	appendComment	客服对评论的回复	reply
第一次发表评论时间	rateDate		

商品评论是顾客对于商品本身体验发表的意见观点；服务评论是顾客对于购买商品过程的服务体验点评，如对物流、客服等评价。商品评论、服务评论、追加评论、客服对评论回复是非结构化文本数据，其他属性都是结构化数据。

表 2 基于 Selenium 的网络评论采集结果之顾客属性集

顾客昵称	displayUserNick	顾客等级	userVipLevel
顾客天猫等级	coustomLever		

顾客昵称是网购商品评论区列表上显示的评论用品昵称。顾客等级对进一步挖掘评论数据的价值具有重要的意义，既可以按评论等级筛选数据，也可以研究不同等级下的用户评论特征。

表 3 基于 Selenium 的网络评论采集结果之交易属性集

交易结束时间	tradeEndTime	选购商品的属性	auctionSku

交易结束时间是指顾客确认收货的时间，可以根据这个属性分析评论对象随时间的变化；选购商品的属性是指顾客购买商品时选择的属性，如购买衣服时

选择的款式、尺寸、颜色等属性，可以通过该属性分析不同颜色、款式的受欢迎程度。

6 结语

本课题研究在充分实验观察了天猫平台页面结构，了解了爬虫和反爬虫的机制后，结合自动化测试框架，构建了天猫平台评论数据抓取的解决方案，并由Java 高级程序设计语言实现了该方案，解决了反爬虫机制封锁的难题，获得了 38 维的高维评论数据。该解决方案具有普遍的适用性，稍做调整即可用于其他互联网平台的数据抓取。本课题研究也有些值得进一步深入研究的地方，如何在抓取的海量评论数据基础上，结合数据挖掘等相关领域的研究进行文本分析，挖掘出评论中的商业价值。

参考文献：

[1] J. Cho. Crawling the web: Discovery and Maintenance of Large-scale Web Data [D]. L.A.: Stanford University, 2001.

[2]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(02):231-237.

[3] 孟庆浩,王晶,沈奇威. 基于 Heritrix 的增量式爬虫设计与实现[J]. 电信技术,2014,(09):97-98+101+99-100.

[4] 宋波，张忠能.　基于系统功能测试的软件自动化测试可行性分析[J].　计算机应用与软件，2005，22(12):31-33.

作者简介：

曹文斌(198-)，男，江西宜黄人，硕士研究生，研究方向为决策理论与决策支持

张科静(1970-)，女，河北鹿泉人，博士，教授，硕士生导师，研究方向为决策分析

收稿日期：2017-12-06　　修稿日期：2018-03-10

Research on Anti-Reptile Technology Based on Automated Testing-Taking Tmall Platform as an Example

CAO Wen-bin，ZHANG Ke-jing

（Glorious Sun School of Business Administration，Donghua University，Shanghai 200051）

Abstract：

With the rapid development of emerging information technologies such as big data, cloud computing and mobile Internet, the dependence of people's work and life on the Internet has been strengthened increasingly. More and more behavioral data and comment data of Internet users have been scattered on the Internet. How to collect data effectively is the premise of analyzing and mining data. Traditional reptiles usually start from a node, and then expand the webpages by links to obtain data blindly and divergently. On the one hand, this method is limited by the technology of anti-reptile technology, on the other hand, the efficiency and quality of data obtained is not satisfied with the requirements. On the basis of researching the webpage structure of Tmall platform deeply, uses the automated testing technology to simulate the way people browse web pages combined with the search engine of E-commerce platform, which effectively avoids the limitations of anti-reptile technology. The correct rate of data collected in batches is more than 96%, meets the actual requirements of scientific, industrial data analysis.

Keywords：

Anti-reptile; Automated Testing; Product Reviews; Selenium