

基于逆向技术的深层网络爬虫与数据分析

邢羽琪, 杨 桢

(云南民族大学数学与计算机科学学院, 云南 昆明 650500)

✉ 20202134050104@ymu.edu.cn; cheng_yang@ymu.edu.cn



摘 要:大数据时代,各行各业对数据采集的需求日益增多,其中使用 JavaScript 加密技术进行数据采集的需求广泛,但也存在不少瓶颈。文章采用 JavaScript 逆向爬虫技术还原参数加密过程,动态构造出某购物网站商品评价的统一资源定位系统(Uniform Resource Locator, URL),实现了指定分类下多商品评价数据的动态采集,为同类加密数据的采集提供了新的思路。使用 SnowNLP[基于 Python 的中文自然语言处理(NLP)库]对采集到的乐高评论数据进行情感分析发现,约 66% 的购买者对商品给出了积极评论;情感分布呈极性,高段集中在 0.8~1.0,低段集中在 0.0~0.2;词云分析显示出购买者群体比较注重商品的快递包装外观。以上结论可为在线商家提升经营管理水平提供参考。

关键词:深层网络爬虫;JavaScript 加密;逆向技术;Ajax;数据挖掘

中图分类号:TP391.1 **文献标志码:**A

Deep Web Crawlers and Data Analysis Based on Reverse Technology

XING Yuqi, YANG Cheng

(School of Mathematics and Computer Science, Yunnan Minzu University, Kunming 650500, China)

✉ 20202134050104@ymu.edu.cn; cheng_yang@ymu.edu.cn

Abstract: In the era of big data, there is an increasing demand for data acquisition from various industries, among which the use of JavaScript encryption technology for data acquisition is widespread, but there are also many bottlenecks. The paper proposes to use JavaScript reverse crawler technology to restore the parameter encryption process and dynamically construct a Uniform Resource Locator (URL) for product evaluation on a shopping website. It realizes the dynamic acquisition of multiple product evaluation data under specified classifications, providing a new approach for the acquisition of similar encrypted data. SnowNLP [Python-based Chinese Natural Language Processing (NLP) library] is used to conduct sentiment analysis on the collected LEGO comment data, and it is found that about 66% of buyers gave positive comments on the product. The distribution of sentiment shows polarity, with high levels concentrated between 0.8 and 1.0, and low levels concentrated between 0.0 and 0.2. Word cloud analysis shows that the buyer group pays more attention to the appearance of the product's express packaging. The above conclusions can provide reference for online sellers to improve their business management.

Key words: deep web crawler; JavaScript encryption; reverse technology; Ajax; data mining

0 引言(Introduction)

互联网经济时代,数据既是基础性资源也是战略性资源,

更是重要的生产力,对各行各业的发展都有着重要影响^[1]。在大数据架构中,网络爬虫技术作为数据获取的重要技术近年来

得以快速发展。另一方面, Ajax(Asynchronous Javascript And XML, 异步的 JavaScript 和 XML)技术在 Web 站点中的使用却增加了数据采集的难度。Ajax 可以实现浏览器与服务器的异步通信, 提升 Web 站点的响应能力。XML 是一种描述、存储、传输数据的标记语言。Ajax 通过使用 XMLHttpRequest 对象从服务器异步请求数据, 并将响应数据以 XML 格式返回。Ajax 技术可以在不刷新整个页面的情况下更新局部内容, 许多网站采用了动态生成页面内容的方式。这使得爬取网页变得更复杂, 网页中的内容可能在用户与页面交互过程中动态加载或改变。爬虫需要解析 JavaScript 代码, 并模拟用户与页面的交互过程, 以获取最终的完整内容。

如今, 大部分 Web 站点都采用 Ajax 动态请求、异步刷新方式生成数据, 使得采取 Python 爬虫爬取静态网页的方法难以直接提取动态网页数据^[2-3]。本文在前人研究方法的基础上, 依据静态网址信息构造动态请求链接, 以爬取某购物网站评论数据的过程为例, 详细描述了构造动态网页 URL 的全过程, 可为 Ajax 动态数据采集技术的研究提供参考。

1 JS 逆向(JS reverse)

JavaScript(通常缩写为 JS)是一种高级、解释型的编程语言, 它的解释器被称为 JavaScript 引擎, 是浏览器的一部分, 广泛用于 HTML 网页, 增强了网页的互动性。通过对 JavaScript 源代码进行混淆加密, 既可以做到网页打开的速度和加密前相差不多, 又可以做到加密结果不可逆, 提高数据的安全性。但是, 对于爬虫技术而言, JS 加密的应用会成为影响数据分析人员采集数据的瓶颈。当前, JS 逆向技术是应对以上瓶颈的有效措施^[4]。JS 逆向技术实质上是根据结果向前推导, 分析构造 URL 所需的参数, 找出后台参数生成原理(即参数加密过程), 动态模拟浏览器 JS 生成结果, 并最终获取目标数据的过程。实施 JS 逆向技术的难点是需要在被改写的源代码中破解加密算法、拆解相关参数。基于此, 本文以某购物网站商品评论页面为例, 采用 JS 逆向技术, 使用 Python 程序还原参数加密过程, 构造完整目标网页的 URL 信息, 实现 Python 环境下的 JS 逆向技术爬取商品评论数据, 同时进行简单的数据挖掘, 使整个数据分析过程完整。

2 评论爬取过程分析 (Analysis of comment crawling process)

爬虫可以分为搜索引擎网络爬虫、基于 Agent 的网络爬虫、迁移的网络爬虫、通用网络爬虫和聚焦爬虫等^[5]。根据要爬取的 Web 页面的存在方式, 又可将爬虫分为表层网络爬虫和深层网络爬虫。表层网络爬虫通常用于爬取以表层网页(即静态网页)为主的 Web 站点。本文的目标购物网站需要用户登录、提交关键词才能动态生成相应页面, 是典型的由深层网页构成的网站, 需要使用深层网络爬虫技术, 动态构造 URL 请求信息, 才能批量爬取指定数量的商品评论信息。基于 Python 的 JS 逆向技术爬虫算法流程图如图 1 所示。

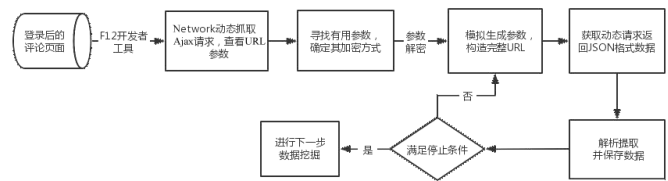


图 1 基于 Python 的 JS 逆向技术爬虫算法流程图

Fig. 1 Flow chart of Python-based JS reverse technology crawler algorithm

2.1 构造 URL

2.1.1 查看 URL 构成

静态网页是指不应用程序而直接或间接制作成 HTML 的网页, 每一个页面都有一个固定的 URL 地址, 可以使用 Python 代码直接获取网页的 URL 和相应的 HTML 信息。动态网页一般使用脚本语言(PHP、ASP 等)将网站内容存储于数据库中, 由于 Python 无法直接获取 URL 动态链接, 因此爬取内容时需要动态构造 URL 链接。

在某购物网站商品评论页面, 利用 F12 开发者工具, 选择 Network 模块, 再筛选 XHR 请求, 刷新网页并不断下拉页, 使开发者工具可以抓取到多个动态 Ajax 请求, 选中其中一个重点查看 URL 的请求头(图 2)。

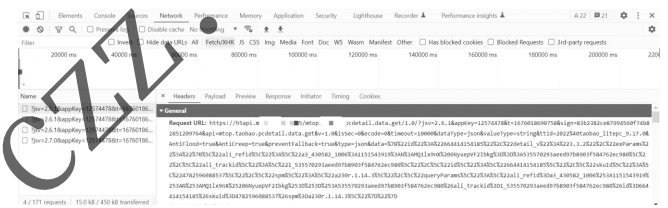


图 2 获取动态 Ajax 请求并查看 URL

Fig. 2 Obtaining dynamic Ajax requests and viewing URLs

2.1.2 确定请求头中的加密参数

在动态 URL 构造过程中, URL 参数的变化信息一般可以在相关静态 URL 及源代码中寻找。通过多次请求分析, 利用静态的 URL 地址的变化构造动态链接, 从而实现数据采集程序自动对动态网页的爬取。

对比多个页面 URL 内容, 最终确定“t”“sign”“data”三个动态变化且对采集有效的参数, 观察多次请求的返回结果, 发现参数“t”为 13 位时间戳, “sign”为加密参数, “data”为商品查询参数。由此, 进入本研究的难点环节, 即逆向破解参数的加密方式以及构造动态 URL。多个页面 URL 参数对比内容如图 3 所示。



图 3 多个页面 URL 参数对比内容

Fig. 3 Comparison of URL parameters in multiple pages

2.1.3 推导参数“sign”加密方式

推导参数“sign”的加密方式,具体步骤如下。

步骤 1:在控制台中使用 length() 函数查看参数“sign”的长度,结果显示 32 位,由此初步推测该网站采用 MD5 加密算法。

步骤 2:使用全局搜索查找“sign”,返回 21 条相关数据,发现返回的 js 文件中多次出现了“sign”,经逐一排查,排除若干干扰项,最终在 mtop.js 文件中追溯到参数“sign”。

步骤 3:查看 mtop.js 文件的源代码,从中筛选出纯参数数据,在其位置打上断点并进行编译。断点查看参数“sign”加密方式如图 4 所示。



图 4 断点查看参数“sign”加密方式

Fig. 4 Breakpoint checking the encryption method of parameter “sign”

步骤 4:查看运行过程得知,构成参数“sign”的信息包括参数“i”“d.token”“g”“c.data”和函数“h”,对其逐一分析发现。参数“i”为实时生成的 13 位时间戳信息;参数“g”由“||”语法得出,“c”是参数对象,因此“g”为请求头中的 appKey,是一个定值;参数“c.data”与参数“g”同理,也是参数对象的“data”字段;参数“d.token”是放在 cookie 信息中的 m_h5_tk 值。对于函数“h”,前文已根据加密结果为 32 位信息推测该网站使用的是 MD5 加密算法,现将上述内容拼接的原始参数使用 MD5 在线加密器计算结果,同时使用断点测试程序查看加密后的结果,对比两者发现结果一致,验证了之前函数“h”为 MD5 加密函数的推测。对比函数“h”与 MD5 在线加密器结果如图 5 所示。



图 5 对比函数“h”与 MD5 在线加密器结果

Fig. 5 Online encryptor results comparison of “h” function and MD5

步骤 5:根据断点运行过程,逐步查看各参数构成方式,据此拼接完整的参数“sign”。

2.1.4 模拟生成 URL

分析 URL 的构成,使用 Python 编写程序,具体为自定义时间戳和随机数,生成并拼接“t”“sign”“data”参数,构造完整的 URL 信息,对整个访问请求进行模拟并获得响应。将上述参数拼接成完整的 URL,用 requests 的 get/post 方法发送请求并接收数据 res = requests.get(url=url, headers=headers), content, 返回得到 JSON 格式数据。

2.2 数据提取

JSON 作为一种轻量级的数据交换格式,结构简单紧凑,是互联网上常用的数据传输与交换格式。JSON 有两种结构形式,一是键值对形式,以“{”开始,以“}”结束,中间部分由 0 个或者多个“,”分隔的键值对构成,是一个无序的“名称/值”对集合;二是数组形式,以“[”开始,以“]”结束,是一个值的有序集合。本次爬取返回的结果是以上两种情况的嵌套,需要通过分析找到所需数据所在层次,即 res['data']['module'] ['reviewVOList']。之后,可利用 Python 爬虫解析库逐条分别提取评论用户、商品详情、评论时间及评论内容等信息。评论页面 JSON 代码的嵌套关系如图 6 所示。



图 6 评论页面 JSON 代码的嵌套关系

Fig. 6 Nested relationship of JSON code for comment pages

2.3 数据保存

使用 pip 安装第三方库,以及对应的数据库软件,可以将结果保存至 MySQL、MongoDB 等数据库中。本文选择自定义 saveCsv() 函数,将爬取结果保存至 csv 文件中,实现该步骤的代码段如下。

```
def saveCsv(datas):  
    with open('result.csv', 'a', newline='') as f:  
        csvwriter = csv.writer(f)  
        csvwriter.writerow([datas['uid'], datas['评论用户'], datas['评论时间'], datas['评论内容']])
```

2.4 结果分析

按照上述流程,本文共获取该网站乐高商品数据 356 条,爬取数据样本示例如表 1 所示。

表 1 爬取数据样本示例

Tab.1 Examples of crawling data samples

商品 ID	评论用户	评论时间	评论内容
638548000000	h * * 1	1 个月前	宝贝不错,物流速度很快,好评
638547618769	露 * * 宝	1 个月前	很好很靠谱
650721272319	麦 * * 5	16 天前	很好 已经拼完了 不愧是经典

2.5 特殊情况

2.5.1 应对反爬

(1)请求头。爬虫可以在程序中添加请求头信息达到伪装成浏览器的目的,从而实现反反爬机制,程序为 `res=requests.get(url=url,headers=headers)`,其中 `url` 为请求网页地址, `headers` 为开发者工具中获取的“User-Agent”请求头信息^[6]。

(2)Cookie。部分需要登录的网站,需要添加 `cookie` 信息。由于 `cookie` 具有生命周期,因此本研究在程序设计上采用分部爬取(20 页信息为一个部分)方式,爬取间隔更换 `cookie` 信息,防止反爬。

(3)Token。Token 实质是访问资源的凭证,一般是用户通过用户名和密码登录成功后,服务器将登录凭证作为数字签名,加密之后得到的字符串即 `token`。查找源代码发现 `token` 就是放在 `cookie` 中的 `m_h5_tk` 值,并且 `token` 值也存在生命周期,需要定期更换。

2.5.2 遵守 robot 协议

在使用爬虫进行数据采集的过程中,应注意遵守对方网站 `robot` 协议的规定,本文在实验过程中充分考虑到这一点,在采集程序加入 `time.sleep(10)` 语句用于将爬虫访问的频率近似到类人水平^[3]。

3 数据挖掘(Data mining)

3.1 数据预处理

由于商品的评论内容是日常口语评论,所以文本存在语言随性、表达多样、噪声大、语言不唯一等情况,进行数据清洗可以让评论文本更加简洁明了。因此,本文对“评论内容”属性列进行去除空格、网址、日期及时间等处理。

3.2 Jieba 分词

中文分词是中文文本处理的一个基础步骤,也是中文人机自然语言交互的基础模块,在进行中文自然语言处理时,通常需要先进行分词。`jieba` 分词支持三种分词模式,即精确模式、全模式、搜索引擎模式^[7]。精确模式是将句子精确地切分,更适用于文本分析;全模式是把句子中所有可以成词的词语扫描出来,较为烦琐;搜索引擎模式是在精确模式的基础上,对长词再次切分,更适用于搜索引擎数据的分词。基于本研究中爬取的数据全部为文本数据,因此采用精确模式 `jieba.lcut()` 对评论内容进行分词,为情感分析环节奠定基础。

3.3 SnowNLP 情感分析

SnowNLP 是一个功能强大的中文文本处理库,它具有中文分词、词性标注、情感分析、文本分类、关键字/摘要提取、TF/IDF、文本相似度等诸多功能^[8]。SnowNLP 利用机器学习和自然语言处理技术,对文本进行情感分析,将其归类为积极、消极或中性等情感类别,并给出相应的情感得分 `emotion`。`emotion` 是一个介于 0 到 1 的实数,其中 0 表示强烈的消极情感,1 表示强烈的积极情感,中间的值表示中性或较弱的情感倾向。

3.3.1 情感得分

使用 SnowNLP 模块对评论进行情感打分,得分示例如表 2 所示。由 SnowNLP 模块计算可知,`emotion` 的平均值约为 0.61,中位数约为 0.81,单纯从评论情感分来看,得出好评多于差评的结果。

表2 SnowNLP 情感得分示例

Tab.2 Examples of SnowNLP sentiment score

评论用户	评论时间	评论内容	emotion
h * * 1	1 个月前	宝贝不错,物流速度很快,好评	0.994 967
露 * * 宝	1 个月前	很好很靠谱	0.795 293
麦 * * 5	16 天前	很好 已经拼完了 不愧是经典	0.997 760
k * * r	26 天前	在这家已经多次购买,包装严实,八角尖尖,安全到货	0.664 583
l * * i	12 天前	压盒	0.500 000

3.3.2 绘制情感分直方图

使用 `matplotlib.pyplot` 库绘制情感分直方图(图 7)。

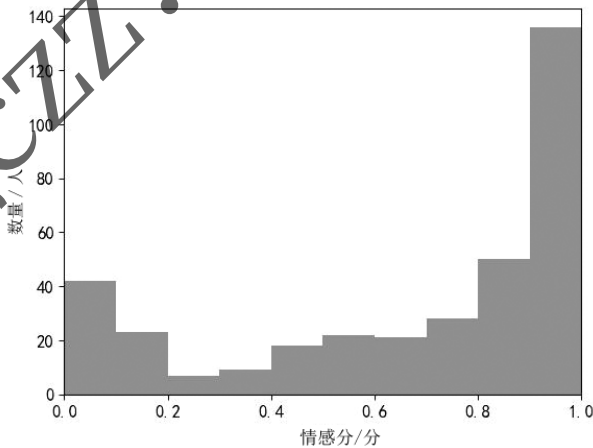


图 7 情感分直方图

Fig. 7 Histogram of sentiment score

由图 7 可以看出,大部分用户的评论情感分较高,介于 0.8~1.0,反映出用户对于乐高销售的整体反响较好,态度积极向上;但也存在部分得分相对较低评论,并且绝大部分得分介于 0.0~0.2,这也对商家起到了警示作用,需要注重用户反馈的信息,及时采取措施纠正或完善。总体来说,评论数据情感分布两级分化较为严重,提醒商家需要调整销售策略,做到扬长避短。

3.3.3 划分积极评论

根据情感评分模型的特点,观察得到当评论为“不错”时,输出得分为 0.861 213;评论为“好”时,输出得分为 0.655 863;评论为“一般”时,输出得分为 0.526 233。根据多条输出结果,判定情感倾向的标准是得分大于等于 0.6 时,可认定为积极态度评论,统计得到积极评论数量为 235 条且占比约为 66%,绘

制其积极评论占比图如图 8 所示。

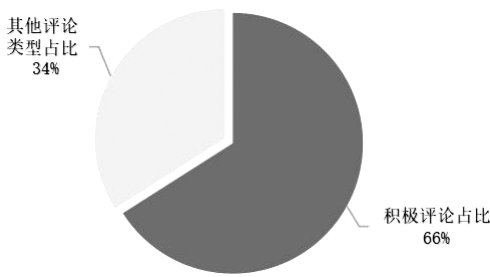


图 8 积极评论占比图

Fig. 8 Proportion of positive comments

根据图 8 得知大部分用户对乐高商品较为满意,可进一步挖掘积极态度的评论,探究用户评论高频词,帮助商家调整销售策略,做到优势最大化。

3.3.4 提取评论高频词

对上述划分为积极态度的评论内容进行分词并按词频统计,提取评论中出现的高频词,并绘制评论高频词云图(图 9)。



图 9 评论高频词云图

Fig. 9 Word cloud diagram of high frequency comments

根据词频展示结果,可以清楚地看到“喜欢”“不错”等字眼占据评论数据的主要地位,“快递”“盒子”“包装”等词汇显示出乐高玩具的用户比较注重快递包装外观,乐高销售商可在这方面进一步做提升,获得更高的用户满意度。

4 结论(Conclusion)

目前,JS 逆向技术俨然成为爬虫领域的重要探索工具:首先通过对网页的结构进行分析,找出请求信息中的加密参数;

其次通过断点技术,寻找有用参数的加密流程;最后模拟浏览器动态构造请求信息获得数据。与传统的爬虫技术相比,整个爬取过程的复杂度和难度都增加了。本文在前人研究的基础上,以某购物网站乐高商品评论数据的爬取过程为例,基于 Python 语言采用 JS 逆向技术,详细描述 JS 逆向技术流程,批量爬取了多个商品的评论数据,并对其结果进行 SnowNLP 情感分析、提取评论高频词,挖掘乐高商品销售可能存在的优势与不足。所使用的爬虫算法及具体实践对使用了 Ajax 和 JS 加密的数据获取有一定的参考作用。

参考文献(References)

[1] 胡美正. 浅析大数据时代的计算机网络安全及防范措施[J]. 电脑知识与技术, 2020,16(36):38-40.

[2] 王佳. 支持 Ajax 技术的主题网络爬虫系统研究与实现[D]. 北京:北京交通大学,2011.

[3] 陈国良,郭修豪. 基于商品评论信息的特征挖掘[J]. 福建电脑,2015,31(5):106-107.

[4] 王朝阳,范伊红,李梦丹,等. Python 环境下的 JavaScript 逆向技术分析[J]. 无线互联科技,2021,18(17):97-98.

[5] 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.

[6] SHRIVASTAVA G K, PATERIYA R K, KAUSHIK P. An efficient focused crawler using LSTM-CNN based deep learning[J]. International Journal of System Assurance Engineering and Management,2023,14(1):391-407.

[7] 韦人予. 中文分词技术研究[J]. 信息与电脑(理论版), 2020,32(10):26-29.

[8] CHEN C X, CHEN J, SHI C. Research on credit evaluation model of online store based on SnowNLP[C]//I-Shou University, University of the West of England, Hongkong Global Scientific Research Association. Proceedings of the 2018 3rd International Conference on Robotics and Automation Engineering. Paris:EDP Sciences,2018:537-540.

作者简介:

邢羽琪(2001-),女,本科生。研究领域:网络爬虫,数据分析。

杨 桢(1980-),女,硕士,副教授。研究领域:数据分析,机器学习,大数据技术。

关于《软件工程》期刊增页的通知

近年来,在主管、主办单位的正确领导下,广大作者、读者的大力支持下,《软件工程》的办刊质量稳步提升,影响力不断扩大,来稿数量与日俱增。但由于版面限制,导致大量优秀稿件积压,影响了论文的时效性与先进性。为了优化资源配置,提升期刊出版服务能力。经辽宁省新闻出版局审核批准,本刊从 2024 年第 1 期起由原来的每期 64 页增加至 80 页。每本定价仍为 10 元。特此通知。欢迎广大作者、读者踊跃投稿与订阅。

《软件工程》编辑部
2023 年 12 月