电子科技 2012 年第 25 卷第 11 期
Electronic Sci. & Tech. /Nov. 15, 2012

基于 Python 的 Web 数据采集技术

齐　鹏，李隐峰，宋玉伟
(西安电子科技大学 电子工程学院，陕西 西安　710126)

摘　要　针对 Web 数据采集技术进行了介绍，分析了 Web 数据采集技术在将非结构化数据转换为结构化数据方面的优势：速度快、准确性高。从 HTTP 协议层分析了 Web 数据抓取的原理，并重点介绍了如何实现基于 Python 的 Web 数据采集方案。Web 数据采集系统可以分为：HTTP 交互和数据解析两个模块。

关键词　Web 数据抓取；屏幕抓取；HTTP 协议；Python；正则表达式；XPath
中图分类号　TP274+.2　　**文献标识码**　A　　**文章编号**　1007－7820(2012)11－118－02

Research on Python-based Web Scraping Technology

QI Peng，LI Yinfeng，SONG Yuwei
(School of Electronic Engineering，Xidian University，Xi'an 710126，China)

Abstract　In this paper web scraping technologies are discussed. The advantages of Web data collection technology for high speed and accuracy conversion of unstructured data into structured data are pointed out. The principles of the web scraping at HTTP level are introduced with emphasis on the technical solutions to Python-based web scraping. Web scraping system consists of two modules: HTTP interaction module and data analysis module.

Keywords　Web scraping; screen scraping; HTTP; Python; regex; XPath

1　Web 数据抓取技术介绍

1.1　Web 数据抓取

Web 数据抓取(Web Scraping)是指从网站上提取信息的一种计算机软件技术。Web 数据抓取程序模拟浏览器的行为，能将任何可以在浏览器上显示的数据提取出来，因此也称为屏幕抓取(Screen Scraping)。Web 数据抓取的最终目的是将非结构化的信息从大量的网页中抽取出来以结构化的方式存储(CSV、JSON、XML、Access、Mssql、Mysql 等)。简而言之，Web 数据采集就是从指定网站抓取所需的非结构化信息数据，分析处理后存储为统一格式的本地数据文件或直接存入本地数据库中。

1.2　Web Scraping 的作用

Internet 是一个巨大的且迅速发展的信息资源。但大多数信息都以无结构的文本形式存在，使得信息归类变得非常困难。在 Web Scraping 出现之前，人们为了归类数据通常会采用手动复制粘贴的方式，这样不但费时费力，而且数据质量得不到保证，效率低。有时遇到海量数据的时候，靠人工整理甚至是无法完成。Web Scraping 是一个使用计算机程序自动从目标网页中摘取某些数据形成统一格式的本地数据的过程，整个过程基本不需要人工干预。其效率较高：

(1) 速度快。抓取程序的数据加载速度要比浏览器快，因为通常情况下浏览器不但要下载基本的HTML数据还需要下载相关的样式表、Java Script 文件、多媒体资源，还要由渲染引擎进行页面排版布局，Java Script引擎还要进行客户端代码执行。而抓取程序只需要下载基本的 HTML 数据即可，这样可缩短数据下载时间。另外，程序的数据提取速度会比人工复制粘贴速度快得多，再结合多线程技术，速度更是人工所无法比拟的。

(2) 准确性高。人工操作会产生信息遗漏或错误的情况，而且纠错难度大。而程序的准确性较高，即便出现问题，纠错也容易，通常只需要修改程序即可。

2　Web Scraping 的原理

Web Scraping 程序在计算机网络通信的传输层，使用 TCP 协议与 Web 服务器进行数据传输，在应用层使用 HTTP 协议与服务器进行数据交互。它与服务器的通信过程和 HTTP 客户端程序浏览器一致。

Web Scraping 程序从功能上可以划分为两大模块：HTTP 交互模块和 HTML 解析模块。对一个网页的抓取过程是：首先 HTTP 交互模块向服务器的 Web 端口发起 TCP 连接，连接建立后，交互模块即可向 Web 服务器发送 HTP 请求报文，当 HTTP 交互模块接

收稿日期：2012-02-24
作者简介：齐鹏(1987—)，男，硕士研究生。研究方向：Web 开发，网络安全。李隐峰(1975—)，男，博士，副教授，硕士生导师。研究方向：Web 开发。宋玉伟(1986—)，女，硕士研究生。研究方向：Web 开发。

收到服务端的应答报文后,进行 HTTP 包拆封,提取其中的 HTML 数据,然后将数据交由 HTML 解析模块进行数据解析和提取,最后解析模块将提取的数据以格式化的形式存储于数据库系统或者是简单的结构化的文本文件(CSV、TSV、XML 等)。整个流程如图 1 所示。

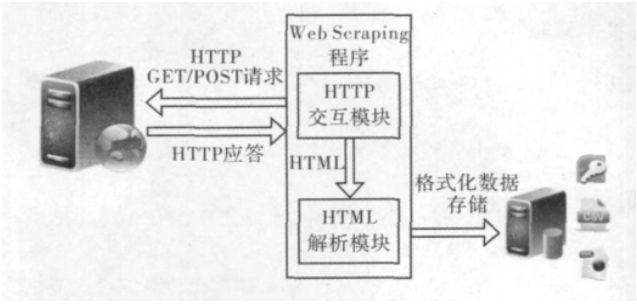

图 1 Web Scraping 的原理

Web Scraping 程序对一个网站的采集过程就是分别对网站内感兴趣的每个页面采集的集合。为得到网站所需要采集的页面地址,需要首选对网站的结构进行分析,总结出页面的规律,例如,网站通常会有一些信息集中的列表页,通过遍历这些列表页即可得到所有详细页面的地址。

对于某些网站在于服务器交互的过程中可能会用到 Cookie,这时就需要抓取程序还能够对 HTTP 报文中的 Cookie 进行管理,例如,当服务端的应答报文中含有 Set - cookie 字段时,要提取 Cookie 数据并在客户端存储或更新;之后发送请求报文时,要将 Cookie 一并发回服务端。

HTML 解析模块负责对 HTML 数据进行提取和规范化处理,然后将数据以结构化的形式存储。

3 基于 Python 的 Web 数据采集方案

Python 是一种面向对象、直译式计算机程序设计语言。其语法简捷而清晰、可读性强、便于维护,并且具有丰富和强大的类库[1]。为 Web Scraping 程序开发提供了便利:可以使用 HTTP 通信模块 urllib2 完成与 Web 服务器的数据交互,使用 cookielib 模块进行 Cookie 管理,使用 re 模块进行文本提取[2],使用 XPath 相关库进行 HTML 解析。总之,Python 提供了 Web Scraping 程序的所有功能模块,利用 Python 可以最少的代码完成功能强大的功能。

3.1 实现 HTTP 交互模块

Python 的 urllib2 模块包含于 Python 的标准库中,它定义了一些类和方法主要用于实现对 HTTP 通信协议的支持。urllib2 支持 HTTP 代理、HTTP 简单认证、跳转、Cookie 等功能[5]。urllib2 模块还支持对 HTTP 请求报文的头和实体进行增改,对 HTTP 应答报文的头和正文进行读取。

如何利用 urllib2 模块进行 HTTP 交互? urllib2. urlopen(url [, data] [, timeout]) 方法提供了最基本的 HTTP 请求构造和 HTTP 应答处理功能。url 参数指示了一个要下载的资源路径。当 data 参数为空时预示着将发出一个 GET 类型的请求,该请求不包含任何实体;当 data 参数为非空时预示着将发出一个 POST 类型的请求,data 的内容即为请求的实体内容。timeout 参数指示了请求超时的时间。

urllib2. urlopen 方法调用的结果有两种情况:

(1) 出现了 HTTP 错误。例如,网络异常或 Web 服务异常造成的请求超时错误;服务端返回 HTTP 错误码。这时 urllib2. urlopen 会抛出一个异常,可以通过捕获不同的异常类型进而判断错误的种类[5]。

(2) 没有出现 HTTP 错误。这时 urllib2. urlopen 返回一个类似文件的对象,通过调用该对象的 read( ) 方法可获取到应答返回的正文内容(HTML) 。如果返回是经过 gzip 压缩过的数据,在这里还要手动进行 gzip 解码。

cookielib 模块也包含于 Python 的标准库中,它主要用于对 Cookie 进行管理[5],urllib2 通过 cookielib 库实现对 Cookie 自动维护。

3.2 实现 HTML 解析模块

通过 HTML 交互模块可以取得网站页面数据,但是此时的数据粗糙,字符编码不确定,结构混乱甚至不符合 XML 规范。所以首先要确定文档的字符编码,通过 < head > 中的 content - type 元得到。然后将其解码成 unicode 类型[3],目的是保证后续数据提取过程中的编码一致性,以及最终数据存储方便。

正则表达式:是指一个用来描述或者匹配一系列符合某个句法规则的字符串的工具。利用正则表达式可以方便地从一堆复杂的文本中找到与规则相匹配的子串,许多程序设计语言都支持利用正则表达式进行字符串操作。在 Python 标准库中 re 是一个用来进行正则表达式相关操作的模块[6]。通过对目标页面结构进行分析,通常能够在感兴趣的字符串周围找到其他有标志性的字符,可以通过这些字符构造出正则表达式,利用 re 模块进行数据提取。

XPath:是一门在 XML 文档中查找信息的语言,用于在 XML 文档中通过元素和属性进行导航。利用 XPath 可以方便地在 HTML 文档中定位感兴趣的节点。lxml 库是 Python 的第三方库,它支持标准的 XPath 规范[7]。

通过正则表达式和 XPath 的结合就可灵活地从 HTML 中提取任何感兴趣的信息。在提取到数据之后

还要对其进行规范化处理,比如将 HTML 转义字符进行反转义、去除冗余的 HTML 标记、去除冗余的空白字符。

3.3 数据结构化存储

结构化数据通常指的是行数据,存储在数据库里,可以用二维表结构来逻辑表达实现的数据。Web Scraping 程序最终输出的数据是结构化的,具体存储于各种数据库系统或文件中。在 unicode 数据进行本地化存储之前必须要先进行字符编码,具体的编码方式可以根据需要选择,一般是由数据最终的应用环境决定的。比如,最终的数据将应用在一个字符编码为 UTF-8 的网站上,那么就要选择以 UTF-8 的编码进行存储。

4 结束语

介绍了 Web 数据抓取技术以及其实现的原理,以及如何利用 Python 进行 Web 数据抓取程序的开发。Web 数据抓取技术已在非结构数据结构化、Web 程序自动化操作、定制搜索引擎爬虫、舆情监控等方面发挥重要作用。同时,为保护好自己的 Web 资源不被别人恶意采集,要做好应对措施,限制网站单个 IP 的并发连接数,可以使用 Ajax 动态加载网页内容或者将应答内容进行加密,将一些敏感的信息以非文本的形式展现,都会给数据采集造成障碍。

参考文献

[1] 赫特兰. Python 基础教程[M]. 2 版. 北京: 人民邮电出版社,2010.
[2] 丘恩. Python 核心编程[M]. 2 版. 北京: 人民邮电出版社,2008.
[3] 鲁特兹. Python 学习手册[M]. 北京: 机械工业出版社,2009.
[4] 桂小林,汪宁波,李文. 基于 XML 的远程教育课件规范化的研究与实现[J]. 电子科技,2010,23(6):129-131.
[5] 刘红梅. 脚本语言在数据采集系统中的应用研究[J]. 电子科技,2009,22(11):72-75.

(上接第 97 页)

3 结束语

随着用电负荷的高速增长,新建变电站的数量将大幅增加,这对调度交换系统的可靠性和稳定性的要求较高。文中针对河池电网现状和双机同组运行的特点,提出在电网中采用调度交换机双机同组调度交换网,为进一步提高电网调度通信系统的防灾减灾能力提供及时的技术参考。

参考文献

[1] 胡平金,苏运东,林苏蓉. 福建电力调度交换网的安全隐患与安全建设思路[J]. 福建电力与电工,2007(1):21-24.
[2] 韦宇宁. 广西电力调度程控交换机组网及应用[J]. 广西电力,2003(2):66-68.
[3] 张立冬,程明. 城市供电网系统可靠性的原因分析[J]. 黑龙江科技信息,2009(26):48-52.
[4] 杨波,孙万蓉,冯战鹏,等. 电力电缆沟道监测系统管理软件的设计[J]. 电子科技,2012,25(5):89-93.
[5] 蔡莹,刘佳明. 陕西中部电网存在问题分析及解决措施建议[J]. 科技资讯,2009(32):11-12.

(上接第 117 页)

4 结束语

针对武器系统软件综合保障工作现阶段开展情况进行了分析,为确保当前装备到部队的武器系统软件综合保障工作的顺利开展,提出了当前软件保障工作的实施要求。同时,作为一门学科,软件保障工作理论的研究和成体系主机制建立还需结合装备部署形势,作进一步的探讨和实践。

参考文献

[1] 中国人民解放军总装备部电子信息基础部. 可靠性维修性保障性术语(GJB 451A-2005)[S]. 北京: 中国人民解放军总装备部,2005.
[2] 国防科工委综合计划部. 装备综合保障通用要求(GJB 3872-99)[S]. 北京: 中国人民解放军总装备部,1999.
[3] 宋华文,耿华芳. 软件密集型装备综合保障[M]. 北京: 国防工业出版社,2011.
[4] 马绍民. 综合保障工程[M]. 北京: 国防工业出版社,1995.
[5] 石柱. 软件质量管理[M]. 北京: 航空工业出版社,2003.
[6] 何国伟. 软件可靠性[M]. 北京: 国防工业出版社,2003.