基于 Python 的网络爬虫系统的设计与实现

李 琳
（河南工业大学信息科学与工程学院,河南 郑州 450001）

摘要:数据的抓取是数据分析工作的基础,没有了数据一些研究分析工作也就无法进行。网络爬虫可以快速抓取互联网各类信息,本文以抓取二手房信息为例,实现基于 Python 的网络爬虫信息系统,完成了目标数据的高效获取。实验结果表明:本程序提供了一种快速获取网页数据的方法,为后续的数据挖掘研究提供支持。

关键词:搜索引擎;Python;网络爬虫
中图分类号:TP393 **文献标识码:**A **文章编号:**1673-1131(2017)09-0026-02

0 引言

网络爬虫是能够自动抓取网页中各类数据的一段程序。网络爬虫通过网页的链接地址来查找网页内容,并直接返回给用户所需要的数据,不需要人工操纵浏览器获取。爬虫是搜索引擎中的重要组成部分,为搜索引擎抓取互联网中的数据。通用的搜索引擎如谷歌、百度等逐渐成为人们访问互联网的入口,但由于其通用性限制,抓取网页时没有针对性,因而也无法对其抓取的结果进行针对特定领域的进一步分析,导致查询结果不够深入和专业化;另外通用的搜索引擎通常会返回一些与用户所寻找的主题无关的结果,造成信息过载。

本文提出的爬虫程序通过模拟登录二手房网址并获取相关数据,并将这些数据保存到本地,方便进一步的数据挖掘与分析。最后本文利用 Python 可视化方法对数据进行简单的数据分析。使用本文爬虫程序可以节省数据分析人员的开发时间,使得他们可以将更多的精力放在数据分析上面,同时也可以对海量数据起到针对性的提取。

1 基于 python 的网络爬虫的设计

1.1 爬虫系统设计需求

设计网络爬虫系统要解决的以下几个问题:

(1)链接网址的提取。首先初始化为一些网址,然后通过网页分析不断抓取新的网址链接。

(2)下载要提取信息的页面。页面上往往有我们需要的信息,如链接、图片、点评等等。

(3)网址的管理。防止网址重复或陷入死循环。

(4)网页内容的分析和管理。提取网页信息,并将其存入数据库或其他数据文件中。

1.2 Python 语言

2.3 现代密码体制的图像加密技术

在信息传输的过程中，彩色图像会转变成为二进制的数据，基于该模式，也可以采用现代的密码体制进行有效改革，主要思想在于通过秘钥,将彩色图像看作成为一个明文,之后运用图 1 所示的原理图,进行信息传递。

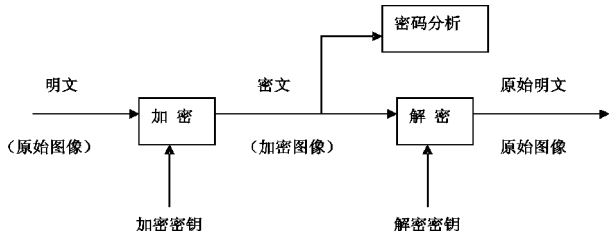

图 1 现代密码体制通信框架图

由于秘钥的不同，现在通用的主要有以下两种不同的现代密码加密体制:第一,对称密钥加密算法。该算法的含义是对秘钥进行反复的推算，无论从正面还是反面进行推算,都能够得到有效的结果。但是有一个前提在于,需要在信息传送之前,制定好秘钥,这是最为安全的性能保证,一旦秘钥出现问题，所有人都能够破解彩色图像中所蕴含的信息;第二,非对称密钥加密算法。该算法的主要思想在于加密和解密的过程是截然不同的,如果知道加密的秘钥,并不能有效推算出解密的秘钥,这样对于加密的信息,就可以进行公开,称之为“公钥”,相对应的就是“私钥”,就是仅仅彩色图像信息拥有者才会知道秘钥。如果运用非对称的秘钥加密算法进行彩色图像加密，就需要对秘钥体制进行保护,防止第三方对解密秘钥的获取或者有意更改信息,对用户进行干扰。现代密码体制的图像加密技术是当今最为普遍也是常用的一种彩色图像加密算法，在很多领域得到了迅速的发展。

3 结语

彩色图像凭借其形象而又生动的表达,在很多重要的场合得到了运用,在受到重视的同时,需要重视彩色图像的传输安全问题。对于彩色图像而言,最为有效的保护措施当属数字加密技术。当前有很多彩色图像的加密技术,传统的加密方式无法满足当前的技术需求,最近几年新兴的算法,例如基于矩阵变换的像素置换、基于秘密分割和共享的图像加密算法等,都能够有效解决彩色图像巨大的数据量以及高相关性等问题,是当今的研究热点,在未来也将有更加广泛的研究和应用。

参考文献:

[1] 朱从旭,孙克辉.对一类超混沌图像加密算法的密码分析与改进[J].物理学报,2012,61(12):12-13.
[2] 刘芳,贾成,冯雁等.图像置乱在数字水印中的应用研究[J].通信技术,2008,41(9):165-167.
[3] 王静,蒋国平.一种超混沌图像加密算法的安全性分析及其改进[J].物理学报,2011,60(6):7-11.
[4] 王英,郑德玲,王振龙.空域彩色图像混沌加密算法[J].计算机辅助设计与图形学学报,2006,18(6).

万方数据
PDF created with pdfFactory Pro trial version www.pdffactory.com

Python 语言是一种功能强大、语法简洁清晰的开源编程语言，几乎能够在目前所有的操作系统上运行；Python 是高效率的完全面向对象的语言，能有效而简单地实现面向对象编程州。Python 解释性语言的本质，再加上其简洁的语法和对动态输入的支持，使得它在大多数操作系统平台上都是一个较为理想的脚本语言，特别适用于快速的应用程序开发。Python 提供了针对网络协议标准库，对网络协议的各个层次进行了抽象封装，程序员就可以集中精力处理程序逻辑。其次，Python 非常擅长处理字节流的各种模式，具有很快的开发速度。

1.3 与爬虫相关的 python 模块

1.3.1 网址管理器

实现网址管理的方法可以分为 3 类：

（1）用内存存储网址，适合数据量较少的情况。将网址存入内存的两个集合，分别表示待爬集合和已爬集合，在 Python 中用 Set()实现。并且 Set()集合本身具有清除重复值的效果；

（2）使用关系数据库，适合永久存储。例如：建立表，其中包括两个字段，分别表示网址以及是否被爬取；

（3）存储在缓存数据库 redis 中，适合存储大数据量的网址。

1.3.2 网页下载器

网页下载器是爬虫程序的主要核心模块。网页的内容一般是 HTML 格式，目前 Python 中支持的网页下载工具有两类：① Python 官方支持的基础模块中的 lib2 包；② request 第三方工具包，功能强大。

1.3.3 网页解析器

网页解析器是对网页内容中进行数据分析的工具。Python 支持的网页解析器有两类方法：一种利用正则表达式可以将整个网页文档当成一个字符串，使用模糊匹配的方式来提取出有价值的数据，虽然直观，但是如果文档比较复杂的话，这种方式非常麻烦，如果一个正则匹配稍有差池，那可能程序就处在永久的循环之中。另一种是根据 Html 网页创建成一个 DOM 树，以树的形式进行各种节点的搜索遍历。DOM 的树形结构根据上下级关系，可以很方便的定位到各个元素。

1.3.4 数据导出

利用 Python 可以将数据导出为 CSV 格式或其它格式文件。导出时应注意数据编码问题，否则导出的文件能会出现乱码。

2 实验设计

下面以爬取“房天下网站”二手房屋的信息为案例，说明爬虫系统的具体实现过程。本文中实验的网址数据不多，因此采用内存存储的方法实现。具体算法步骤如下.

第 1 步：预处理。打开要保存的 csv 文件，写入标题行内容；同时写入爬取的网址。

第 2 步：分析并提取网页中的数据。步骤：①利用 Requests 对象打开页面的网址；②利用 BeautifulSoup 解析网页；③分析网页中要提取的各部分的数据格式。

图 1 为要抓取的“房天下”的页面，右半部分是对应的 html 结构文档：例如：HTML 页面中房屋的“标题信息”的标记顺序为 dd-p-a。对应的具体代码：titles = soup.select('dd > p.title > a')。

图 1 Html 网页及对应的结构文档

第 3 步：保存数据。步骤：①将提取出来的数据保存到字典对象中。②向 csv 文件中保存字典数据。

3 实验结果

通过实验获取“房天下”中二手房屋的信息，其运行结果如图 2 所示。实验采集了 12705 条房屋信息，为后期的数据分析提供了有力的依据。

图 2 已抓取的二手房信息

4 数据分析

对已抓取的数据进行面积和单价的数据可视化分析，得到图。从图中可以看出面积在 75-100 之间的房屋和 125-150 之间房屋的单价大多较高，平均价格也比较高，数据为购房者提供了一定的参考价值。

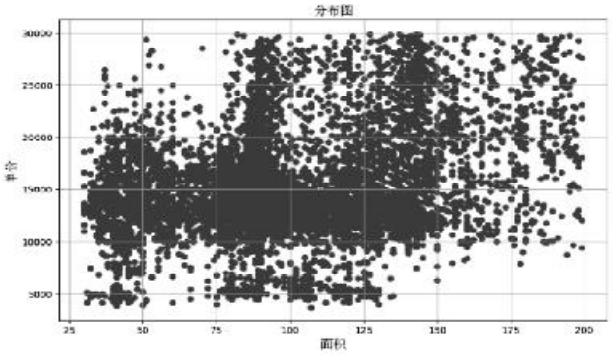

图 3 单价-面积数据分布图

5 结语

总之，网络爬虫技术具有较高的应用价值， 通过抓取数据，可以挖掘出更有价值的信息；同时 Python 语言功能强大，提供了各种软件工具包支持，利用 Python 可以方便有效实现对 Web 数据信息的抓取。

参考文献：

[1] Mark Lutz.Learning Python[M].北京:机械工业出版总社,2009.
[2] 于娟,刘强.主题网络爬虫研究综述叨.计算机工程与科学,2015,02:231-237.

万方数据
PDF created with pdfFactory Pro trial version www.pdffactory.com