**分类号：** (顶格，小四号宋体)

**学校代码：11460**

**学 号：14552608**

**南京晓庄学院本科生毕业**

**设计**

**Web自动化嗅探下载器的设计与实现**

**Design and implementation of Web automatic sniffer Downloader**

**所属学院： 信息工程学院**

**学生姓名： 阙志兴**

**指导教师： 明子成**

**职称：**

**研究起止日期：二○一七 年 十 月至 二○一八 年 五 月**

**学位论文的独创性声明**

本人郑重声明：

1.坚持以“求实、创新”的科学精神从事研究工作。

2.本论文是我个人在导师指导下进行的研究工作和取得的研究成果。

3.本论文中除引文外，所有实验、数据和有关材料均是真实的。

4.本论文中除引文和致谢的内容外，不包含其他人或其它机构已经发表或撰写过的研究成果。

5.其他同志对本研究所做的贡献均已在论文中作了声明并表示了谢意。

作者签名：

2018年5月12日

摘 要

在今天的互联网时代，数据正在爆炸式的产生。从互联网获取资源的种类、数量不断增多的同时，为获取真正要找的资源而花费的时间也越来越多。

本文的目的就在于设计出能够自动化地从web上嗅探并下载满足特定需求的应用程序。该程序采用当前对数据处理最为方便的python编程语言，能够针对特定格式、主题、数量等要求对网络资源进行嗅探并下载，节约人们获取特定资源的时间。与此同时，本文还针对如何对特定文件类型进行嗅探下载的方法加以描述，对读者添加其他类型资源的嗅探下载方法提供借鉴作用。

**【关键词】**：资源嗅探； 资源下载； 自动化； web； python

Abstract

In today's Internet era, data is exploding. With the increasing number and types of resources acquired from the Internet, more and more time is being spent to get the resources that are really needed.

The purpose of this paper is to design and automate the sniffing and downloading of applications that meet specific needs from web. The program uses the most convenient python programming language, which is most convenient for data processing. It can sniff and download the network resources for specific format, topic, quantity and so on, so as to save time for people to obtain specific resources. At the same time, this paper also describes how to sniff downloads for specific file types, and provides a reference for readers to add other types of resources to sniffer downloading.

【**Keywords**】**:** **Resource sniffing; Resource download; Automation; Web; Python**

**目录**

[摘 要 3](#_Toc513667193)

[Abstract 4](#_Toc513667194)

[1 绪 论 6](#_Toc513667195)

[1.1 课题背景及意义 6](#_Toc513667196)

[1.2 研究现状 6](#_Toc513667197)

[1.3 研究内容 7](#_Toc513667198)

[1.4 论文结构 7](#_Toc513667199)

[2 开发环境与相关技术 8](#_Toc513667200)

[2.1 Windows10操作系统 8](#_Toc513667201)

[2.2 Mysql数据库 8](#_Toc513667202)

[2.3 Python编程语言 8](#_Toc513667203)

[2.3.1 Python语言的特点 8](#_Toc513667204)

[3 系统分析 8](#_Toc513667205)

[4 系统设计 9](#_Toc513667206)

[4.1 系统总体设计 9](#_Toc513667207)

[4.2 数据库设计 9](#_Toc513667208)

[5 系统实现 9](#_Toc513667209)

[6 系统测试 9](#_Toc513667210)

[6.1 系统测试目的 9](#_Toc513667211)

[6.2 系统功能测试 9](#_Toc513667212)

[6.3 测试结果 9](#_Toc513667213)

[7 总结 9](#_Toc513667214)

[参 考 文 献 10](#_Toc513667215)

[致谢 10](#_Toc513667216)

1. 绪 论
   1. 课题背景及意义

自信息化时代以来，人们可以从互联网上获取各种各样的信息资源，有视频、音频、图片、文本等不同资源类型，每种资源类型又有多种编码格式，造成对特定数据进行搜寻的各种困难与不便。在互联网上获取需要的特定资源所需要的步骤是非常多的：首先需要打开搜索引擎；然后选择提供资源的特定网站；其次在网站内对我们需要的资源进行简单描述并搜索；接着筛选出需要下载或浏览的具体条目；最后点击下载或进入详细描述页面。在如上步骤中，得到人们所需结果之前的各种步骤所花费的时间可能是最后一个步骤花费时间的好几倍，所以人们需要为了仅仅下载或查看某一两个资源而花费成倍的时间在搜寻工作上。

21世纪的生活节奏快到前所未有。时间紧张，生活的步伐不断加快，人们都在抱怨自己有多么忙碌，所以选择节约时间的生活及工作方式成为必然。选择外卖把出门吃饭的时间省下来，选择网购把出门购物的时间省下来、选择视频会议把开会聚集的时间省下来。本课题的研究意义就在于能够减少人们从互联网上搜寻特定资源所花的时间，而且减少的时间与总时间的比例随着下载资源数量增多而增加，更加适应当前数据爆炸式增长的时代。

* 1. 研究现状

业界对于嗅探器有两个研究方向：

其一为在数据链路层上拦截网络适配器收到的数据封包，将它们首先传递给某些能够直接访问数据链路层的软件，逆向解析还原帧的内容。然后嗅探着就根据自己的需要挑选自己觉得有用的信息，完成对网络的嗅探。

此类[嗅探](https://baike.baidu.com/item/%E5%97%85%E6%8E%A2" \t "_blank)器最初由 Network General 推出，由 Network Associates 所有。后来，Network Associates 决定另开辟一个嗅探器产品单元，该单元组成一家私有企业并重新命名为 Network General，如今嗅探器已成为 Network General 公司的一种特征产品商标，由于专业人士的普遍使用，嗅探器广泛应用于所有能够捕获和分析[网络流量](https://baike.baidu.com/item/%E7%BD%91%E7%BB%9C%E6%B5%81%E9%87%8F/7489548)的产品。

其二为在应用层基于Http协议，根据url向互联网各个站点发出请求，再对站点返回的网页数据进行解析，嗅探出网页数据中不同类型的文件，然后再根据用户预先设置的筛选条件对资源进行自动化下载。

此类嗅探器有许多成型产品，包括用户安装之后使用的独立软件，也包括再浏览网页时实时嗅探当期网页资源的与浏览器插件。

* 1. 研究内容

本课题主要研究的是一个能依据用户特定需求进行自动化嗅探下载资源的应用软件。用户可以从图形化交互界面选择或输入需要嗅探下载的资源的筛选条件。在用户输入数据之后程序能够自动化地嗅探各站点的网页资源。

在完成对某类数据类型进行自动化嗅探下载后，整理出后续添加其他数据类型嗅探下载的思路与方法。

* 1. 论文结构

本文结构安排如下：

1.绪论：简要介绍本文的课题意义，并且对本文的主要工作以及主要论文结构进行一个简单的介绍说明。

2.开发环境与相关技术：简要介绍要完成本系统所需要的开发环境、相关技术。

3.系统分析：分析系统的功能需求、性能需求、安全性、可靠性。

4.系统设计： 在需求分析的基础上完成各模块的功能设计，数据库设计等。

5.系统实现： 给出核心功能的算法设计，并给出后续扩展的思路。

6.总结：针对完成整个项目以来所遇到的问题以及收获做一个简要的叙述，突出完成本毕设的意义。

1. 开发环境与相关技术
   1. Windows10操作系统

Windows 10系统成为了智能手机、PC、平板、Xbox One、物联网和其他各种办公设备的心脏，使设备之间提供无缝的操作体验。Windows 10操作系统在易用性和安全性方面有了极大的提升，除了针对云服务、智能移动设备、自然人机交互等新技术进行融台外，还对固态硬盘、生物识别、高分辨率屏幕等硬件进行了优化完善与支持。从技术角度来讲，Windows 10操作系统是一款优秀的消费级别操作系统。

* 1. SQLite数据库

SQLite是一个轻量级、跨平台的关系型数据库。有如下优点与缺点：

* + 1. 技术上的优点和特性

1. 轻量级

SQLite和C/S模式的数据库软件不同，它是进程内的数据库引擎，因此不存在数据库的客户端和服务器。使用SQLite一般只需要带上它的一个动态 库，就可以享受它的全部功能。而且那个动态库的尺寸也挺小，以版本3.6.11为例，Windows下487KB、Linux下347KB。

2. 绿色软件

SQLite的核心引擎本身不依赖第三方的软件，使用它也不需要“安装”。所以在部署的时候能够省去不少麻烦。

3. 单一文件

所谓的“单一文件”，就是数据库中所有的信息（比如表、视图、触发器、等）都包含在一个文件内。这个文件可以copy到其它目录或其它机器上，也照用不误。

4. 跨平台/可移植性

如果光支持主流操作系统，那就没啥好吹嘘的了。除了主流操作系统，SQLite还支持了很多冷门的操作系统。我个人比较感兴趣的是它对很多嵌入式系统（比如Android、Windows Mobile、Symbin、Palm、VxWorks等）的支持。

5. 内存数据库（in-memory database）

内存越来越便宜，很多普通PC都开始以GB为单位来衡量内存。这时候，SQLite的内存数据库特性就越发显得好用。

SQLite的API不区分当前操作的数据库是在内存还是在文件（对于存储介质是透明的）。所以如果磁盘I/O有可能成为瓶颈的话，可以考虑切换 为内存方式。切换的时候，操作SQLite的代码基本不用大改，只要在开始时把文件Load到内存，结束时把内存的数据库Dump回文件就OK了。

* + 1. 技术上的缺点和不足

1.并发访问的锁机制

SQLite在并发（包括多进程和多线程）读写方面的性能一直不太理想。数据库可能会被写操作独占，从而导致其它读写操作阻塞或出错。

2.SQL标准支持不全

在它的官方网站上，具体列举了不支持哪些SQL92标准。特别是不支持外键约束。

3.网络文件系统（以下简称NFS）

有时候需要访问其它机器上的SQLite数据库文件，就会把数据库文件放置到网络共享目录上。这时候当SQLite文件放置于NFS时，在并发读写的情况下可能会出问题（比如数据损坏）。原因据说是由于某些NFS的文件锁实现上有Bug。

* 1. Python编程语言

Python是一种解释型、交互式、纯面向对象的脚本程序设计语言,它结合了多种不同语言最经典的设计原则和思想。自1990年Guido van Rossum 发明Python以来,它的简单、清晰和优雅吸引了广泛的关注,Python 社区以每两年翻番的速度迅速扩张。

Python是纯粹的自由软件， 源代码和解释器CPython遵循 GPL(GNU General Public License)协议。Python语法简洁清晰，特色之一是强制用空白符(white space)作为语句缩进。 Python具有丰富和强大的库。它常被昵称为胶水语言，能够把用其他语言制作的各种模块（尤其是C/C++）很轻松地联结在一起。常见的一种应用情形是，使用Python快速生成程序的原型（有时甚至是程序的最终界面），然后对其中有特别要求的部分，用更合适的语言改写，比如3D游戏中的图形渲染模块，性能要求特别高，就可以用C/C++重写，而后封装为Python可以调用的扩展类库。

* + 1. Python语言的特点

相比其他任何程序设计语言, Python允许软件设计者更关注于问题本身而不是实现细节, 其主要特性有:

(l)面向对象：Python 是一种纯面向对象语言, 支持面向对象的所有机制, 这使它可以应用于大型软件项目开发。

(2) 虚拟机机制：Python代码编译为字节码在虚拟机上执行,这种类似Java的机制使Python具有平台无关性, 并获得比其他脚本语言更快的执行效率。

(3) 内置高级数据类型：Python中的元组、列表、数组、字典等高级数据类型具有极强的表达能力, 几乎任意复杂的对象在Python中均可以以最简单的方式进行描述。

(4) 交互式开发方式和单元测试：类似Foxpro 和 Matlab ,Python 提供了内置交互式窗口, 可以允许开发者边尝试边书写代码, 极大地提高了软件开发效率。另外单元测试机制在每个类代码完成后均可以单独执行、测试。这两种开发方式使得”搭积木” 式的软件开发过程成为现实。

(5) 动态执行特性：类似早期Dbase系统的宏代替、Python允许动态构造程序代码并执行, 使得软件中的重复代码得到简化。

(6) 高度可扩展性Python：语言本身只实现了一个小内核,其强大功能体现在大量的扩展模块。计算机科学领域几乎所有的已知算法均可找到Python语言扩展。不仅c / c + + 编写的程序可以方便地改写成Python扩展模块,而且Python自身也可以被嵌人到口C/C++项目中。

(7)平台无关性：Python字节码可以不经修改跨平台运行在所有主流操作系统上, 甚至包括Pocket PC 、Palm等嵌入式系统。

1. 系统分析
   1. 软件功能需求分析

为了实现减少人们从互联网上搜寻特定资源所花的时间这一重要意义，本嗅探软件应当满足下列需求。

1. 能让用户在较友好的界面下选择文件类型需求，如图片、视频等。

2. 能让用户针对特定文件类型进行选择更加细致的筛选条件，如资源主题、数量等。

3. 能让用户一次输入得到多个输出，实现自动化嗅探下载。

* 1. 软件性能需求分析

人们原先从互联网上获取资源需要如下步骤：

首先需要进入浏览器打开搜索引擎(T1)；然后选择特定网站；其次在网站进行搜索；接着筛选出具体条目；最后点击下载。

。。。

1. 系统设计
   1. 系统总体设计
   2. 数据库设计

由于SQLite的单一文件特性，所以它支持列的亲和类型概念。任何列仍然可以存储任何类型的数据，当数据插入时，该字段的数据将会优先采用亲缘类型作为该值的存储方式。为了读者可以移植到其他类型的关系数据库，特意对数据类型和长度也作出规划。

根据软件中的资源类型实体设计出的数据表如下：

图片实体数据表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 列名 | 数据类型 | 长度 | 亲和类型 | 允许为空 | 说明 |
| Id | int | 20 | integer | 否 | 图片序列号 |
| Name | varchar | 50 | text | 是 | 图片名 |
| Title | varchar | 50 | text | 否 | 图片主题 |
| Width | int | 10 | text | 是 | 图片宽度 |
| Height | int | 10 | text | 是 | 图片高度 |
| Format | varchar | 10 | text | 是 | 图片格式（后缀名） |
| Description | varchar | 255 | text | 是 | 图片描述 |
| Descrurl | varchar | 255 | text | 是 | 图片详细描述地址 |
| Thumburl | varchar | 255 | text | 是 | 图片预览地址 |
| Downurl | varchar | 255 | text | 是 | 图片下载地址 |

数据表 4‑1

论文实体数据表

1. 系统实现

（我来写）

1. 系统测试
   1. 系统测试目的
   2. 系统功能测试

（我来写）

* 1. 测试结果

（我来写）

1. 总结

参 考 文 献

[1]

致谢