基于 Python 的图书馆业务报表自动生成研究

张怡华

（成都信息工程大学，四川成都 610225）

摘要：Python 丰富的标准库提供了强大的网络处理和文本分析功能，实现了业务数据自动的提取并形成报表。本文采用 B/S 模式的图书馆业务管理系统进行分析，研究 Python 的发展简介，分析其优点和缺点。接着针对 Python 的图书馆业务报表系统，阐述其工作流程和主要技术，分析报表的生成和实现。

关键词：Python；图书馆；业务报表

中图分类号：D920.4 **文献标识码：**A **文章编号：**2096-4609（2018）48-0218-002

一、Python 的发展简介

（一）Python 的发展历史

最初的 Python 由 Guido 设计，Guido 希望有一种语言，这种语言能够像 C 语言那样，能够全面调用计算机的功能接口，又可以像 shell 那样，可以轻松的编程。1991 年，第一个 Python 编译器（同时也是解释器）诞生。它是用 C 语言实现的，并能够调用 C 库（.so 文件）。从 Python 语法很多来自 C，但又受到 ABC 语言的强烈影响。来自 ABC 语言的一些规定直到今天还富有争议，比如强制缩进。但这些语法规定让 Python 容易读。另一方面，Python 聪明的选择服从一些惯例（特别是 C 语言的惯例）。比如使用等号赋值，使用 def 来定义函数。Guido 认为，如果“常识”上确立的东西，没有必要过度纠结。

（二）Python 的优点

1. 可拓展性强

Python 可以在多个层次上拓展。从高层上，你可以引入 .py 文件。在底层，你可以引用 C 语言的库。Python 程序员可以快速的使用 Python 写 .py 文件作为拓展模块。但当性能是考虑的重要因素时，Python 程序员可以深入底层，写 C 程序，编译为 .so 文件引入到 Python 中使用。Python 就好像是使用钢构建房一样，先规定好大的框架。而程序员可以在此框架下相当自由的拓展或更改。

2. 对象与过程均支持

面向过程和面对对象是一种编程思想，不能说某某语言是不是面对对象或是面向过程，而是某某语言是否支持面对对象或面向过程。回归主题，python 可以支持面对对象，但也可以支持面向过程，即使不支持面向对象的语言，比如 c 语言，也可以用面对对象的思想写程序。你可以理解面对对象为“模块化”，恰巧 python 可以做到这一点，自己编写的函数文件可以用 import 引用模块，即使不使用 class 定义类，也可以实现面对对象的思想。

3. 可扩展性与嵌入性强

在嵌入式开发领域中开发语言以 C/C++ 为主，如今基于 Python 的 MicroPython 已经涉入到该领域中，MicroPython 是一位叫 Damien George 的工程师，基于 ANSI C(C 语言标准），然后在语法上又遵循了 Python 的规范，主要是为了能在嵌入式硬件上（这里特指微控制器级别）更易于的实现对底层的操作。

4. 代码规范可读性强

Python 不算最好的胶水，至少 Lua 做胶水就比 Python 好，API 好用。但 Python 再算上强大的标准库、数据结构、友好的转义后，而且简单易懂，做 Web 很方便。

（三）Python 的缺点

1. 强制缩进

Python 和自然语言十分相近：编写容易，维护容易，开发迅速。但是其强制性的要求语法缩进。python 里面都是强制缩进，所以代码结构清晰，保证你过再长的时间来看，对整个程序的结构都是一清二楚。

2. 适用受局限

因为很多编程都是不严肃的场合，比如用完就扔了，解决临时问题，特定环境，测试性质，业余玩票，插件扩展等等，属于快餐文化。而在特定的严肃场合，python 适用受局限。

3. 架构选择分散

python 的瓶颈在性能，但是按照现在硬件的水平，这个问题越来越不是主要矛盾了，除非你是对性能要求极其苛刻的任务，大多数情况下我们用 python 都能对付，另外相比较语言本身，代码的优化更值关注。

二、基于 Python 的图书馆业务报表系统

（一）系统工作流程

1. 1interlib 系统登录

图书馆业务报表生成系统发展迅速，现用的基本都是以 Interlib 为代表的全新的第三代图书馆自动化编目管理系统。Interlib 采用基于 web 和 Internet 的 B/S 模式，用户端不需要安装软件即可实现图书馆业务工作。因此本文选取 Interlib 进行操作，用 Interlib 系统实现了总分馆模式。

图书馆业务报表生成的工作流程是：Interlib 采用 B/S 体系架构，先用浏览器登录，登陆时输入正确的工作人员用户名和密码验证。Interlib 系统中通过区域图书馆群的联合、协调采购，能够提前合理配置。浏览器提交涵盖 URL、cookies 和 post 表单信息响应后，返回包含数据的 HTML 文件。登录后可以实时处理读者办证、图书借还等流通工作，而且也可以获取阅览人数、外借人数、外借册次等统计信息。

2. 模拟登录

依次选择系统——系统参数设置——点击【新增】按钮。

```
import datetime
import urllib.request
import http.cookiejar
import re
workbook = xlsxwriter.Workbook(‘chart.xlsx’)
worksheet = workbook.add_worksheet()
chart = workbook.add_chart({‘type’: ‘column’})
```

3. 报表实现

（1）Requests：有 http 请求，需要用到

Requests；（2）BeautifulSoup

需要简单的从网页上爬去一些数据；（3）xlrd，xlsxwriterexcel 的读写操作，通常用 xlrd 读，用 xlsxwriter 写（效率高），生成报表后即可使用邮件处理程序发送到指定联系人。

（二）主要技术

1. 基于 Python 的动态网页快照

用 mongoDB 来存储网页的快照，首先用 Python 的 urllib.request 弄个爬虫，定期爬取要监控网页，和存储在库的快照比较。比较直接用 Python 自己带的 difflib 库，因为现在网页基本上都是动态生成的，内容没有变化，但是有日期时间，或推荐之类的都会变化。

2.Email 发送

（1）class email.message.Message：__getitem__，__setitem__ 实现 obj[key] 形式的访问。Msg.attach(playload): 向当前 Msg 添加 playload。Msg.set_playload(playload): 把整个 Msg 对象的邮件体设成 playload。Msg.add_header(_name, _value, **_params): 添加邮件头字段。

（2）class email.mime.base.MIMEBase(_maintype, _subtype, **_params) 所有 MIME 类的基类，是 email.message.Message 类的子类。

（3）class email.mime.multipart.MIMEMultipart() 在 3.0 版本的 email 模块 (Python 2.3-Python 2.5) 中，这个类位于 email.MIMEMultipart.MIMEMultipart。这个类是 MIMEBase 的直接子类，用来生成包含多个部分的邮件体的 MIME 对象。

（4）class email.mime.text.MIMEText(_text) 使用字符串 _text 来生成 MIME 对象的主体文本。

3. 任务计划

Linux 任务计划

一次性任务计划 at 命令

服务“atd”（”service atd status” 查看服务是否启动，通过”chkconfig –level 35 atd on”从 3 和 5 级别启动”atd”）

#at 17:30 2012 或者 #at 17:30

>./report.sh

> shutdown –h now

> 按 Crtl+d 结束

#at –l #atq 列出 at 计划

#at –d 计划编号 #atrm 计划编号 删除任务计划

注意 :1、任务计划的编号只会增长。

4. 手机邮箱

手机邮箱运营商为移动手机用户提供的新一代移动办公产品，它将邮件主动推送到用户手机上。手机邮箱是基于国际最先进的 pushmail 技术至上的业务，可以支持多种终端，尤其是手机终端的访问。

三、报表实现的技术方案

（一）获取 HTML 文件

1. 1fiddler 进行抓包

我们通过 fiddler 进行抓包，分析发现，系统存储各种统计数据的地址相同。在大型网站的架构中，大多需要多个子域名，这些子域名可能是单独用于缓存静态资源的，也可能是专门负责媒体资源的，或者是专门负责数据统计的（如 pingback）。

2. 提供查询条件

我们可以通过提供查询条件和统计项目来获取存储数据的 HTML 文件，<pre-name=”code”class=”plain”deep=”7”>// 查询

```
$scope.ishow=function(ages){// 年龄
Var range_age=$scope.age;//21-30
if(range_age==undefined||range_age==”“){
Return true;}
Var arr_age= range_age.split(“-”);
Var min=arr_age[0];
Var max=arr_age[1];
if(ages<min||ages>max){
return false;
}else{return true;}
```

3. 调用 read 方法进行读取并解码

编码是将一组字符转换为一个字节序列的过程 , 解码是反向的操作，利用 re 模块提供的正则表达式首先将上述两种字符串从 HTML 文件中提取出来

（二）提取统计数据

Python 的 pandas 包提供一种 group 格式，即 dict(字典格式)，然后利用 describe 方法输出统计结果 pandas 是 pypi 提供的众多包之一，其中提供了大量的统计方法。针对图书馆日常工作中需要定期总结、汇报业务数据的问题我们利找到相应业务统计功能 , 获得存储数据的页面 , 利用正则表达式提取数据。

（三）汇总统计数据

有时候我们在统计相同 key 值的时候，希望把所有相同 key 的条目添加到以 key 为键的一个字典中，然后再进行各种操作，我们将数据填入之后，相当于进行快速分组，然后遍历每个组就可以统计一些我们需要的数据。data 是我们的格式数据，使用 zip 后进行快速键值转换，然后可以使用 max，min 之类函数进行数据操作。数据格式就是 data，我们想要对 name 或者 uid 进行排序我们就是用代码中的方法。在进行分组前要首先对数据进行排序处理，排序字段根据实际要求来选择即将处理的数据。

四、结语

本文利用 Python 提供的丰富的网络处理和文本分析标准库，解决图书馆日常工作中需要定期总结、汇报业务数据的问题，提高了业务工作的效率。Python 可对代码因地制宜适合自身需求，在后续工作中，我们采用多样化模式的图书馆业务管理系统进行研究。

【作者简介】张怡华（1978–），女，助理研究员，本科，研究方向为数字图书馆技术与管理。

【参考文献】

[1] 辛海滨 . 基于 Python 的图书馆业务报表自动生成研究 [J]. 电脑知识与技术 : 学术交流 ,2016,12(9X):72–74.

[2] 王朝阳 . 基于 Python 的图书信息系统的设计与实现 [D]. 吉林大学 , 016.

[3] 李琳 . 基于 Python 的网络爬虫系统的设计与实现 [J]. 信息通信 ,2017(9):26–27.

[4] 高瑞华 . 基于 APK 文件抓取系统的匹配模块设计 [J]. 电子设计工程 , 2016, 24(3):47–49.

[5] 温晓明 ,WENXiao–ming. 基于 Python 的电子资源可用性检测方案 [J]. 中华医学图书情报杂志 ,2013, 2(1):68–71.

【相关链接】

图书馆，是搜集、整理、收藏图书资料以供人阅览、参考的机构，早在公元前 3000 年就出现了图书馆，图书馆有保存人类文化遗产、开发信息资源、参与社会教育等职能。据《在辞典中出现的"图书馆"》说，"图书馆"一词最初在日本的文献中出现是 1877 年的事 ; 而最早在我国文献中出现，当推《教育世界》第 62 期中所刊出的一篇《拟设简便图书馆说》，时为 1894 年。中国最早的省级图书馆为 1904 年创办的湖北省图书馆。早在公元前 3000 年时，巴比伦的神庙中就收藏有刻在胶泥板上的各类记载。最早的藏书地点是希腊神庙的藏书之所和附属于希腊哲学书院 (公元前 4 世纪) 的藏书之所。