ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.12, No.27, September 2016
E-mail: xsjl@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

基于Python的图书馆业务报表自动生成研究

辛海滨

（济南市图书馆，山东 济南 250017）

摘要：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，该文利用Python实现了业务报表的自动生成。Python丰富的标准库提供了强大的网络处理和文本分析功能。该文通过分析报表生成的基本工作原理，利用Python实现模拟登陆、获取HTML文件、提取数据，最终汇总形成报表。

关键词：图书馆；业务统计；Python；自动生成

中图分类号：TP315 文献标识码：A 文章编号：1009-3044(2016)27-0072-03

DOI:10.14004/j.cnki.ckt.2016.3534

信息技术快速发展的今天，国内大部分图书馆都已经配备了计算机系统，使图书馆基本业务实现了自动化，给图书馆工作带来了极大的便利。但除这些基本业务活动外，各个业务部门在日常工作中还要定期总结、汇报业务数据（如周报、月报等），使管理层能够及时掌握业务部门的运行情况。以报刊借阅室为例，业务数据包括阅览人次、流通人数、期刊外借册次以及新刊记到种数、册数等。这些数据存在于图书馆业务管理系统的不同功能模块下。当获取某一项业务数据时，我们需要逐层展开系统功能模块，找到相应业务统计功能，选择各项查询条件，如开始日期、结束日期和部门代码等，等待系统输出统计数据，并将之记录下来。其他业务数据的统计也遵循同样的流程。最后我们将获取的所有业务数据汇总到一个文件中，行成业务报表。这些工作具有重复、繁琐的特点，消耗了工作人员相当大的精力去细心、耐心对待。Python是一种解释型编程语言，提供了功能强大的用于网络处理的标准库。可以利用Python提供的这些库登陆系统、获取数据，实现业务报表的自动生成。这可以减少人为失误，提升业务数据统计的准确度、提高工作人员的工作效率。

1 基本工作原理

当前存在多种图书馆业务管理系统被不同图书馆应用，Interlib采用基于web和Internet的B/S模式，实现了图书馆业务在线管理，具有代表性，因此本文选取Interlib进行操作。生成业务报表的基本工作原理如图1所示。我们首先通过浏览器进行系统登录，登录成功后，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并汇总形成报表。

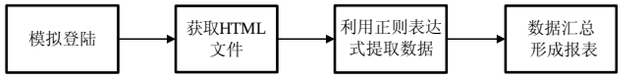

图1 工作原理

登录时，在浏览器中打开interlib系统登录页面，工作人员输入用户名、密码等登录信息，提交给服务器。服务器响应，返回包含数据的网页。从浏览器层面看，浏览器提交包含URL、cookies和post表单等必要信息的请求，服务器响应，返回包含数据的HTML文件。其中URL（Uniform Resource Locator）是统一资源定位符，是资源在互联网上位置和访问方法的一种简洁表示；cookies是存储在本地的文本文件，用于辨识用户和进行会话跟踪；post表单是调用POST方法时提交的用户请求表单等信息。登录后，从系统获取阅览人数、外借人数、外借册次等统计信息时，浏览器发出包含各种查询条件的请求，服务器响应，返回包含统计数据的HTML文件。HTML文件是一种由标签和内容组成的文本文件，浏览器通过标签来显示文件中的内容。我们利用正则表达式从返回的HTML文件中获取需要的数据。最后，将各项业务统计数据汇总成报表。

本文利用fiddler抓取网页进行分析，获取URL和需要提交的post表单等信息。Interlib系统服务器的IP地址假定为“192.168.0.1”。工作人员用户名和密码分别以“myusername”和“mypassword”来表示。

2 模拟登录

首先导入开发过程中所使用的库：

```
import datetime
import urllib.request
import http.cookiejar
import re
```

datetime模块提供了处理日期和时间的类，用于处理程序中的日期、时间数据；re模块是Python用于实现正则表达式的库，我们可以通过re模块提供的功能来提取HTML文件中的特定内容。http.cookiejar模块定义了自动处理HTTP cookies的类。Http.request模块定义了一系列通过复杂方式打开URL的类和方法。我们创建CookieJar对象来存储http cookies，绑定http处理器。代码如下：

```
cj=http.cookiejar.CookieJar()
pro=urllib.request.HTTPCookieProcessor(cj)
opener=urllib.request.build_opener(pro)
```

工作人员输入用户名和密码登录Interlib系统，如图2a所示。我们将登录过程通过fiddler抓包进行分析，如图2b所示。从用线圈出的部分可以看出登录URL为http://192.168.0.1/in-

收稿日期：2016-08-10

terlib/common/Login。post表单数据编码成字节字符串后为：b'cmdACT=opLOGIN&furl=maxMain.jsp&passwd=34819d7beeabb9260a5c854bc85b3e44&competno=999libcodeadmin&loginid=myusername&passwd_in=mypassword'。我们定义函数login()进行系统登录，通过提供URL和post表单打开网页。主要代码如下：

```
def login():
    url_login='http://192.168.0.1/interlib/common/Login'
    postDict_login=b'cmdACT=opLOGIN&furl=maxMain.jsp&passwd=34819d7beeabb9260a5c854bc85b3e44&competno=999libcodeadmin&loginid=myusername&passwd_in=mypassword'
    data_login=opener.open(url_login,postData_login)
```

通过cookies处理器构建的opener能够自动管理cookies，不必关心其实现细节。

图2a　登录Interlib系统

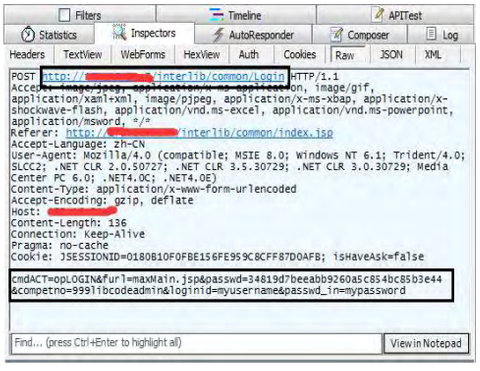

图2b　通过fiddler抓包进行分析

3 报表实现

我们调用login函数成功登录系统后，打开系统相应业务数据统计功能，通过查询条件获取存储统计数据的HTML文件。分析HTML文件结构，利用正则表达式提取我们需要的数据。

3.1 获取HTML文件

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，即URL为http://192.168.0.1/interlib/report/StatisServlet。各业务统计功能提交的Post表单包含相同的5个数据项，分别为cmdACT、mod、xsl、whereSQL和statisId，其中三个数据项cmdACT、mod和xsl的内容针对各种业务统计没有发生变化，数据项whereSQL表示查询条件，数据项statisId用于标识业务统计项目。如此，我们就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，主要代码如下：

```
def getData(whereSQL='',statisId=''):
    url_getData='http://172.16.31.7/interlib/report/StatisServlet'
    postDict_getData={
        'cmdACT':'doStatis',
        'mod':'oneXSL',
        'xsl':'',
        'whereSQL':whereSQL,
        'statisId':statisId}
    postData_getData=urllib.parse.urlencode(postDict_getData).encode()
    data_getData=opener.open(url_getData,postData_getData).read().decode()
    return data_getData
```

其中postData_getData=urllib.parse.urlencode(postDict_getData).encode()是应用urllib.parse模块的urlencode函数将post表单数据进行编码。将它同URL传递给opener返回一个HTTPResponse对象，调用read方法进行读取并解码，得到我们需要的HTML文件，如图3b所示。本文以文献借还册次统计中统计文献借阅情况进行说明，图3a为文献借还册次统计结果页面，图3b为其相应的HTML源文件。

文献借还册次统计

[redacted]　借还时间：2016-03-02 00:00:00~2016-03-03 00:00:00　统计类型：--借阅统计--

(索书号)	册数
A|马列主义	0
B|哲　学	1
C|社科总论	15
D|政治法律	15
E|军　事	0
F|经　济	7
G|文科教体	23
H|语　言	0
I|文　学	9
J|艺　术	5
K|历史地理	0
N|自科总论	3
O|数　理　化	0
P|天文地球	0
Q|生物科学	0
R|医药卫生	7
S|农业科学	4
T|工业技术	3
U|交通运输	0
V|航空航天	0
X|环境劳保	0
Z|综合图书	15
其它	0
合计	107

图3a 文献借还册次统计结果

图3b 文献借还册次统计结果HTML源文件

3.2 提取统计数据

通过分析图3a、3b发现，HTML源代码中不含合计数据，只包含各大类的文献借阅册次数据。这就要求我们提取出各大

类的数据，并将之加和得到需要的数据。在各种业务统计中，数据都存储在<VAL1 code="16">和<VAL2 code="166">这种类型的格式中，一般VAL1存储人数、册数等类型数据，VAL2存储人次、册次等类型的数据。图3b所示的文献借还册次统计中只包含VAL2标签，即该HTML文件中只包含文献借阅册次数据。我们定义函数getvalue，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，然后从得到的字符串中提取出数据。代码如下：

```
def getvalue(data=''):
    return_ls=[]
    zs_re=r'<VAL1 code="\d{1,}">'
    zs_exp=re.compile(zs_re)
    zs_list=zs_exp.findall(data)

    cs_re=r'<VAL2 code="\d{1,}">'
    cs_exp=re.compile(cs_re)
    cs_list=cs_exp.findall(data)

    sum_zs=0
    sum_cs=0
    for row1 in zs_list:
        num1=re.sub(r'<VAL1 code="|">','',row1)
        sum_zs+=int(num1)
    for row2 in cs_list:
        num2=re.sub(r'<VAL2 code="|">','',row2)
        sum_cs+=int(num2)

    return_ls.append(sum_zs)
    return_ls.append(sum_cs)

    return return_ls
```

在此功能中，我们定义列表return_ls存储加总的统计数据，第一项存储人数、册数等类型数据，第二项存储人次、册次等类型数据。代码中，列表zs_list存储HTML文件中所有的<VAL1 code="数字">字符串，cs_list存储<VAL2 code="数字">字符串。

3.3 汇总统计数据

在生成业务报表过程中，需要将各项业务数据逐一提取出来，最后汇总到一个文件中。获取某项统计数据时，输入开始时间和结束时间并通过给statisId赋值确定业务统计项目，即可获取该数据。文献借阅数据统计的主要代码如下：

```
def jhcctj(s_date,e_date):
    whereSQL="(log_cir.regtime between TO_DATE('" + s_date + "','YYYY-MM-DD') and (TO_DATE('" + e_date + "','YYYY-MM-DD')+1)) and log_cir.data2 in (select rdid from reader where rdlib ='分馆代码') and holding.orglocal='部门代码' and log_cir.libcode ='分馆代码' and log_cir.userid in('部门工作人员') and (log_cir.logtype='30001' or log_cir.logtype='30003' or log_cir.logtype='30007' or log_cir.logtype='30009' or log_cir.logtype='30050' or log_cir.logtype='30051')"
    statisId='bookLoanBookTimesSta'
    data_jhcctj=getData(whereSQL,statisId)
    ls_jhcctj=getvalue(data_jhcctj)
```

其中s_date和e_date两个参数分别为开始日期和结束日期，将之传入whereSQL，形成查询条件。statisId='bookLoanBookTimesSta'表示要进行的业务统计项目是文献借还册次统计。调用getData函数获取对应的HTML文件，再对该HTML文件调用getvalue函数获取数据。我们从图2b中看出，该HTML文件中只包含VAL2标签，即该统计中只有文献借阅册次而没有借阅册数数据。我们统计2016年3月2日至3日的文献借阅册次数据，运行程序，得到列表ls_jhcctj的值为[0，107]，即2016年3月2日至3日文献借阅册次为107册次。

本文应用tkinter设计了一个简单的界面，将部门各项统计数据提取、汇总，如图4所示。

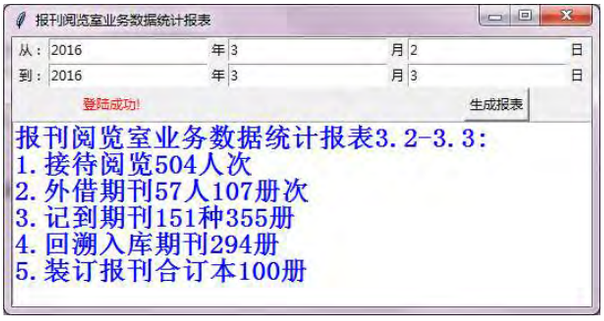

图4 汇总统计数据

4 结束语

本文利用Python提供的丰富的网络处理和文本分析标准库，实现了业务数据自动的提取并形成报表。这使工作人员摆脱了重复繁琐的数据提取工作，能够把更多的精力投入到业务工作中。Python简单易学、功能强大，工作人员在具体工作中可对代码进行修改以适应自身需求，如提取某个工作人员的业务数据。本文不足之处在于该业务报表只是针对采用B/S模式的图书馆业务管理系统实现。在后续工作中，我们将对采用其他模式的图书馆业务管理系统进行研究。

参考文献：

[1] 刘艳平，俞海英，戎沁. Python模拟登录网站并抓取网页的方法[J]. 微电脑应用. 2015(1): 58-60.
[2] 李东来，宛玲，金武刚. 公共图书馆信息技术应用[M]. 北京：北京师范大学出版社，2013(1).
[3] 梁勇. Python语言程序设计(英文版)[M]. 北京：机械工业出版社，2013.