3.4 爬取网站复杂数据

3.4.1 Web 服务器网站

我们进一步把前面的 Web 网站的 mysql.html,python.htm,java.htm 丰富其中的内容，并加上图形：

(1) mysql.htm

```
<h3>MySQL 数据库</h3>
<div>
MySQL 是一个关系型数据库管理系统，由瑞典 MySQL AB 公司开发，目前属于 Oracle 旗下产品。MySQL 是最流行的关系型数据库管理系统之一，在 WEB 应用方面，MySQL 是最好的 RDBMS (Relational Database Management System，关系数据库管理系统) 应用软件。
</div>
<div>
<img src="mysql.jpg"    />
</div>
<a href="books.htm">Home</a>
```

(2) java.htm

```
<h3>Java 程序设计</h3>
<div>
    Java 是一门面向对象编程语言，不仅吸收了 C++语言的各种优点，还摒弃了 C++里难以理解的多继承、指针等概念，因此 Java 语言具有功能强大和简单易用两个特征。Java 语言作为静态面向对象编程语言的代表，极好地实现了面向对象理论，允许程序员以优雅的思维方式进行复杂的编程
</div>
<div>
    <img src="java.jpg" />
</div>
<a href="books.htm">Home</a>
```

(3) python.htm

```
<h3>Python 程序设计</h3>
<div>
    Python （英国发音：/ˈpaɪθən/ 美国发音：/ˈpaɪθɑːn/），是一种面向对象的解释型计算机程序设计语言，由荷兰人 Guido van Rossum 于 1989 年发明，第一个公开发行版发行于 1991 年。
</div>
<div>
<img src="python.jpg" />
</div>
<a href="books.htm">Home</a>
```

3.4.2 爬取网站的复杂数据

好了，我们来爬取网站中的 myslq,python,java 的简介与图像。我们看到简介在网页的第一个 中，图像在<img>中，而且只有这 3 个网页有这样的特征，于是我们设计客户端程序如下：

```
from bs4 import BeautifulSoup
import urllib.request

def spider(url):
    global urls
    if url not in urls:
        urls.append(url)
        try:
            data=urllib.request.urlopen(url)
            data=data.read()
            data=data.decode()
            soup=BeautifulSoup(data,"lxml")
            print(soup.find("h3").text)
            divs=soup.select("div")
            imgs=soup.select("img")
            if len(divs)>0 and len(imgs)>0:
                print(divs[0].text)
                url=start_url+"/"+imgs[0]["src"]
                urllib.request.urlretrieve(url,"downloaded "+imgs[0]["src"])
                print("downloaded ",imgs[0]["src"])
            links=soup.select("a")
            for link in links:
                href=link["href"]
                url=start_url+"/"+href
                spider(url)
        except Exception as err:
            print(err)

start_url="http://127.0.0.1:5000"
urls=[]
spider(start_url)
print("The End")
```

程序结果：

```
计算机
数据库
MySQL 数据库
```

MySQL 是一个关系型数据库管理系统，由瑞典 MySQL AB 公司开发，目前属于 Oracle 旗下产品。MySQL 是最流行的关系型数据库管理系统之一，在 WEB 应用方面，MySQL 是最好的 RDBMS (Relational Database Management System，关系数据库管理系统) 应用软件。

downloaded mysql.jpg
计算机
程序设计
Python 程序设计

Python （英国发音：/ˈpaɪθən/ 美国发音：/ˈpaɪθɑːn/），是一种面向对象的解释型计算机程序设计语言，由荷兰人 Guido van Rossum 于 1989 年发明，第一个公开发行版发行于 1991 年。

downloaded python.jpg
Java 程序设计

Java 是一门面向对象编程语言，不仅吸收了 C++语言的各种优点，还摒弃了 C++里难以理解的多继承、指针等概念，因此 Java 语言具有功能强大和简单易用两个特征。Java 语言作为静态面向对象编程语言的代表，极好地实现了面向对象理论，允许程序员以优雅的思维方式进行复杂的编程

downloaded java.jpg
计算机网络
The End

程序执行完毕后还看到下载了 3 个文件："downloaded mysql.jpg"、"downloaded python.jpg"、"downloaded java.jpg"

其中程序部分：

```
if len(divs) > 0 and len(imgs) > 0:
    print(divs[0].text)
    url = start_url + "/" + imgs[0]["src"]
    urllib.request.urlretrieve(url, "downloaded " + imgs[0]["src"])
    print("downloaded ", imgs[0]["src"])
```

是判断这个 url 页面是否有 与<img>，如果有的就获取第一个 的文字，下载第一个<img>的图像。

3.4.3 爬取程序的改进

1、服务器程序

由于我们的 web 网站时本地的，因此下载图像非常快，而实际应用中 Web 网站是远程的一个服务器，由于网络原因可能下载会比较慢。为了模拟这个过程，我们修改服务器程序如下：

```
import flask
import os
import random
import time

app=flask.Flask(__name__)
```

```
def getFile(fileName):
    data=b""
    if os.path.exists(fileName):
        fobj=open(fileName,"rb")
        data=fobj.read()
        fobj.close()
        #随机等待 1-10 秒
        time.sleep(random.randint(1,10))
    return data

@app.route("/")
def index():
    return getFile("books.htm")

@app.route("/<section>")
def process(section):
    data=""
    if section!="":
        data=getFile(section)
    return data

if __name__=="__main__":
    app.run()
```

这个程序在每次返回一个网页或者图像的函数 getFile 中都随机等待了 1-10 秒，这个过程十分类似网络条件较差的情景，即访问任何一个网页或者图像都有 1-10 秒的延迟。

2、客户端程序

从目前的程序来看这个程序在下载一个图像时是等待的，如果这个图像很大，那么下载时间很长，程序就必须一直等待，其它网页就无法继续访问下去了，即卡死在一个网页的图像下载处。为了避免这个问题，一般可以对程序做以下改进：

(1) 设置 urllib.request 下载图像的时间，如果超过一定时间还没有完成下载就放弃；

(2) 设置下载过程是一个与主线程不同的子线程，子线程完成下载任务，不影响主线程继续访问别的网页。

改进后的客户端程序如下：

```
from bs4 import BeautifulSoup
import urllib.request
import threading

def download(url,fileName):
    try:
        #设置下载时间最长 100 秒
        data=urllib.request.urlopen(url,timeout=100)
        data=data.read()
```

```
            fobj=open("downloaded "+fileName,"wb")
            fobj.write(data)
            fobj.close()
            print("downloaded ", fileName)
        except Exception as err:
            print(err)

def spider(url):
    global urls
    if url not in urls:
        urls.append(url)
        try:
            data=urllib.request.urlopen(url)
            data=data.read()
            data=data.decode()
            soup=BeautifulSoup(data,"lxml")
            print(soup.find("h3").text)
            links=soup.select("a")
            divs=soup.select("div")
            imgs=soup.select("img")
            if len(divs)>0 and len(imgs)>0:
                note=divs[0].text
                print(note)
                url=start_url+"/"+imgs[0]["src"]

                #启动一个下载线程下载图像
                T=threading.Thread(target=download,args=(url,imgs[0]["src"]))
                T.setDaemon(False)
                T.start()
                threads.append(T)

            for link in links:
                href=link["href"]
                url=start_url+"/"+href
                spider(url)
        except Exception as err:
            print(err)

start_url="http://127.0.0.1:5000"
urls=[]
threads=[]
spider(start_url)
#等待所有线程执行完毕
for t in threads:
```

```
    t.join()
print("The End")
```

```
程序一次执行的结果：
计算机
数据库
MySQL 数据库
```

MySQL 是一个关系型数据库管理系统，由瑞典 MySQL AB 公司开发，目前属于 Oracle 旗下产品。MySQL 是最流行的关系型数据库管理系统之一，在 WEB 应用方面，MySQL 是最好的 RDBMS (Relational Database Management System，关系数据库管理系统) 应用软件。

```
计算机
downloaded  mysql.jpg
程序设计
Python 程序设计
```

Python （英国发音：/ˈpaɪθən/ 美国发音：/ˈpaɪθɑːn/），是一种面向对象的解释型计算机程序设计语言，由荷兰人 Guido van Rossum 于 1989 年发明，第一个公开发行版发行于 1991 年。

```
Java 程序设计
```

Java 是一门面向对象编程语言，不仅吸收了 C++语言的各种优点，还摒弃了 C++里难以理解的多继承、指针等概念，因此 Java 语言具有功能强大和简单易用两个特征。Java 语言作为静态面向对象编程语言的代表，极好地实现了面向对象理论，允许程序员以优雅的思维方式进行复杂的编程[1

```
downloaded  python.jpg
计算机网络
downloaded  java.jpg
The End
```

从结果看到访问 python.htm 网页后没有及时完成 python.jpg 的下载，python.jpg 是在访问 java.htm 网页后才完成下载的，这就是多线程的过程。