Knownsec

Spider

程序文档

Wenyu Zhang
2015/5/3

目录

1 设计部分.....2
1.1 说明.....2
1.2 Spider 需求.....2
1.3 Spider 设计.....2
2 使用部分.....4
2.1 Spider 程序.....4
2.2 Spider 使用.....5
3 遇到问题.....6

1 设计部分

1.1 说明

文档仅对 Spider 程序简单说明，未按软件工程思想撰写，适用于程序的使用者及开发者。

1.2 Spider 需求

编写页面抓取程序，程序能够抓取给定网站的页面，保存网站的 url 列表，并将抓到的每个页面单独保存成文件。

说明：

[1] 使用 gevent 或多线程；

[2] 能够设定参数，如深度、最大页面数等；

[3] 记录必要的日志；

[4] 页面存成文件后，给定原来的 url 能够迅速找到相应的文件；

1.3 Spider 设计

根据需求对 Spider 程序功能进行如下设计：

[1] 采用多线程方式抓取页面，利用 threading 模块实现。

[2] 页面抓取采用广度遍历算法。

[3] 采用命令行设定程序必要的参数，利用 optparse 模块实现。

[4] 程序实时记录日志信息，通过文件（logging.conf）配置日志输出级别、位置及名称，主要利用 logging 模块实现。

[5] 抓取的页面保存在 save 目录中，每个 url 对应一个文件，以 url 的 md5 值作为文件名称。

[6] 所有 url 及 md5 值均保存在内存字典中，程序退出时该字典序列化保存在 spider. cpickle 中。

[7] 通过 url 快速查找对应的文件，将 spider. cpickle 反序列化加载到内存中，

计算 url 的 md5 值找到对于的文件名称。

[8] 提供调试选项（--debug）。

1.3 Spider 实现

程序使用的模块包括 os、sys、Queue、time、threading、logging、logging.config、optparse、urllib、urllib2、socket、hashlib、gzip、cPickle 等（详见源码）。

2 使用部分

2.1 Spider 程序

程序的目录结构如下：

./spider

./search.py
./spider.py
./conf
./conf/logging.conf
./conf/config.xml
./threads
./threads/__init__.py
./threads/threads.py
./modules
./modules/__init__.py
./modules/md5.py
./modules/serialize.py
./modules/debug.py
./save
./doc
./log
./crawl
./crawl/__init__.py
./crawl/crawler.py
./crawl/url.py
./README.md

注：

序号	模块名	备注
1	spider.py	程序入口，提供命令行解析，配置等功能
2	search.py	提供查找 url 对应的文件功能
3	logging.conf	日志的配置文件
4	threads.py	抓取页面工作线程
5	serialize.py	提供序列化和反序列化功能
6	debug.py	输出 debug 信息
7	md5.py	计算 md5 值
8	crawler.py	抓取页面
9	url.py	匹配页面中的 url
10	/log、/save	保存日志和页面文件的目录，自动生成

2.2 Spider 使用

程序执行：

```
root@ubuntu:~/spider# ls
conf  crawl  doc  log  modules  README.md  save  search.py  spider.py  threads
root@ubuntu:~/spider#
root@ubuntu:~/spider#
root@ubuntu:~/spider# python spider.py -u http://www.jxnu.edu.cn -d -p 100
```

debug 模式：

```
root@ubuntu:~/spider#
root@ubuntu:~/spider#
root@ubuntu:~/spider# python spider.py -u http://www.jxnu.edu.cn --debug
```

按 url 查找文件：

```
root@ubuntu:~/spider# ./search.py

Please input url or Quit:
http://news.jxnu.edu.cn
```

3 遇到问题

[1] 提取页面 url 链接的规则不是很通用。

[2] 页面抓取线程会出现卡死现象，使得程序无法退出，需要添加处理机制（超时）。

[3] 线程池相对于现在的多线程开销会少一些，后面会改进。

[4] 时间原因尚未进行完整的测试。