测

试

报

告

|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 作者 | 版本描述 | 完成日期 |
| 1.0 | 安志斌 | 基本实现对定向网站的数据抓取 | 2016.06.12 |

目录

[第一节 测试目的 3](#_Toc453517694)

[第二节 反爬测试 3](#_Toc453517695)

[2.1 下载延迟测试 3](#_Toc453517696)

[2.2 代理IP测试 4](#_Toc453517697)

# 第一节 测试目的

在正式爬取目标网站之前，往往需要进行一定程度的压力测试，弄清楚该网站服务器的反爬机制等访问规则。本报告针对目标网站http://www.qincai.net

进行了两项测试：下载延迟时间测试和代理IP访问次数限制。

# 第二节 反爬测试

## 2.1 下载延迟测试

如果网站访问频率过快，可能导致服务器过载，引起网站拥有者关注。所以要设置适当的下载延迟时间。在scrapy框架下，下载延迟时间是指下载器在完成本次下载任务到下次下载任务开始执行前需要等待的时间，其设置位于setting.py文件中DOWNLOAD\_DELAY变量，表示单位为秒。

在测试中，我们设定一系列DOWNLOAD\_DELAY：5,3,1,0.8,0.4,0.1,0，需要注意的是，假设DOWNLOAD\_DELAY=5,并非两次下载间隔就一定为5秒，实际上间隔为位于0.5\*5秒到1.5\*5秒之间的一个随机数，这是由于scrapy框架默认对该延迟进行了随机处理，即DONDOMIZE\_DOWNALOAD\_DELAY。

不同测试条件下，我们打开爬取到数据库中的记录如下图2.1，通过记录数与耗时的比值粗略获得抓取速度。结论：DOWNLOAD\_DELAY值越小，抓取速度越快。

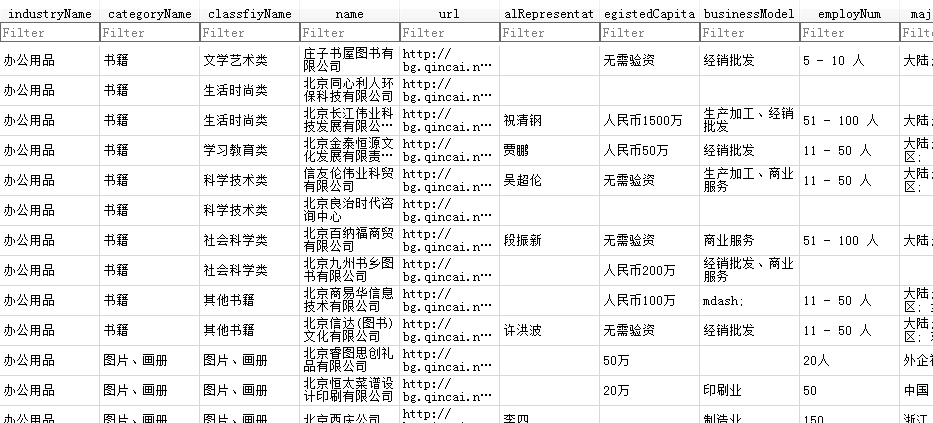


图 2.1 本地数据库记录a

当然，下载延迟测试并非只为获得一个定性的结论，在不同测试条件下，我们还需要获得一个“下限值”，即DOWNLOAD\_DELAY值到多少的时候，抓取效率由于受到服务器的访问限制变得异常低时对应值。针对<http://www.qincai.net>网站，初步分析，DOWNLOADER\_DELAY可以取默认的零值，这是由于使用了代理IP后自然地其爬取速度就下降了一个层次。

## 2.2 代理IP测试

大规模爬取网站数据最为重要的一环就是IP代理设置，而对其进行测试有利于优化爬取算法，提高信息采集的效率。本次测试的逻辑如下：选取某一稳定代理IP，高强度访问目标网站，观察访问次数达到多少时，该IP被服务器禁止访问。

我们选取某一高质量匿名代理101.96.11.30:8080进行测试，测试发现，该网站几乎没有访问次数限制，图2.2为程序获得的运行日志统计情况，图2.3为抓取到数据库中的全部记录数，访问次数接近22000次。

所以，下一步大规模爬取逻辑设计应尽可能考虑让同一成功的IP尽多的再次访问网页以提高效率。

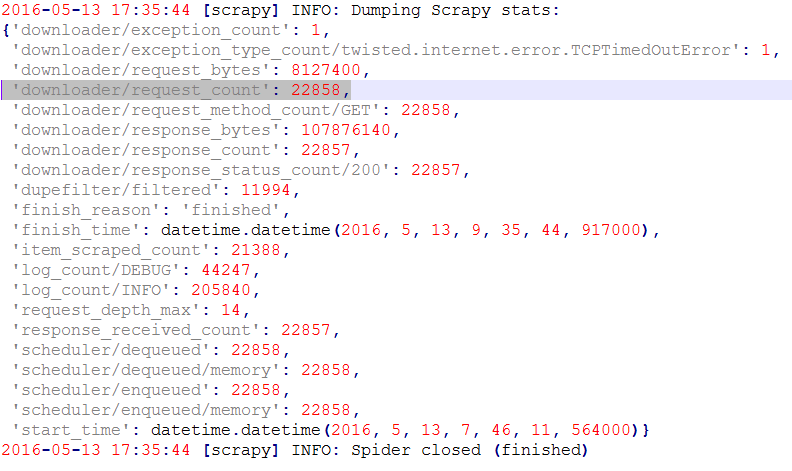


图 2.2 程序某次运行统计

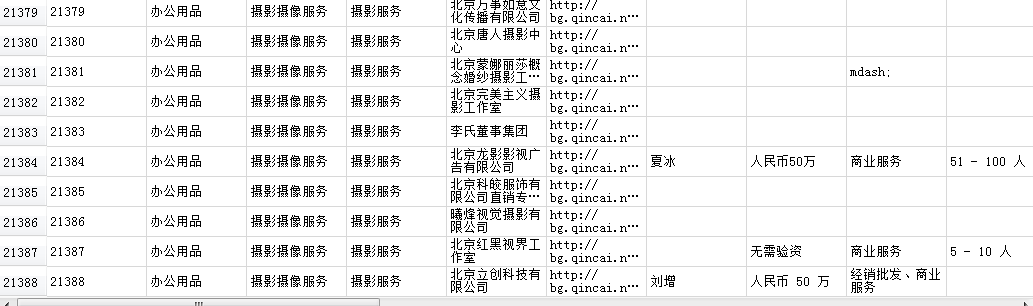


图 2.3 本地数据库记录b