# 改进与扩展

## 改进内容

1.1.1 Scrapy分布式爬虫实现

Scrapy是基于Twisted异步框架的单机爬虫，目前只能在一台机器上进行爬取，而且受限于Python的GIL，在多核CPU上也只能利用到一个CPU核心。所以，我们决定将scrapy改造成一个多机分布式爬虫，实现多台机器对同一个爬取任务的共同完成。

1.1.2 Scrapy两级过滤器改进与实现

Scrapy采用Python内置的集合（set）实现内存过滤器，同时提供文件过滤器让用户选择。两者并不排斥，文件过滤器的设置是为了可以停止爬虫（结束爬虫进程），然后在之后的某一个时间内重新接着上次的爬取¬¬继续时，可以直接从文件中读取已经爬取的url。实际运行期间的过滤都是在内存过滤器（set集合实现的过滤器）进行的。

默认的Python集合需要存储网页本身的url，同时为了实现管理集合以及其他一些针对集合的操作，Python需要保存的内容除了URL之外还有许多，因此，当爬取大量网页的时候占用内存会比较大。而分析发现，我们实际上只需要判断一个url是否已经出现过，而至于url是什么或者何时出现并不在乎。其次，我们也不会对已经爬取的url进行索引等操作，所以Python内置的set集合本身的许多操作相对来说是冗余的。

结合我们决定要实现的Scrapy分布式爬虫，我们决定结合布隆过滤器实现两级过滤器。单机爬虫首先在本机的布隆过滤器中对新生成的url进行过滤，过滤失败则在继续分布式的过滤器中进行过滤。

1.1.3 反爬改进

由于要使用Scrapy进行测试，不得不大量爬取某个网站的内容。然而，在Scrapy提供的下载中间件机制，可以在请求队列与下载请求之间做一些动作。Scrapy本身也提供了一个ProxyMiddleware，但是它只能使用固定的IP地址。由于固定的IP地址可能遭到IP被禁用的情况，这么一来，你就不能使用本地IP进行相应的爬虫工作了。所以，此时就需要进行反爬虫的实现，使用到的两种基本策略为：伪装user-agent和免费代理

1.1.4 网页展示

由前端网页展示后台分布式爬虫抓取的结果。

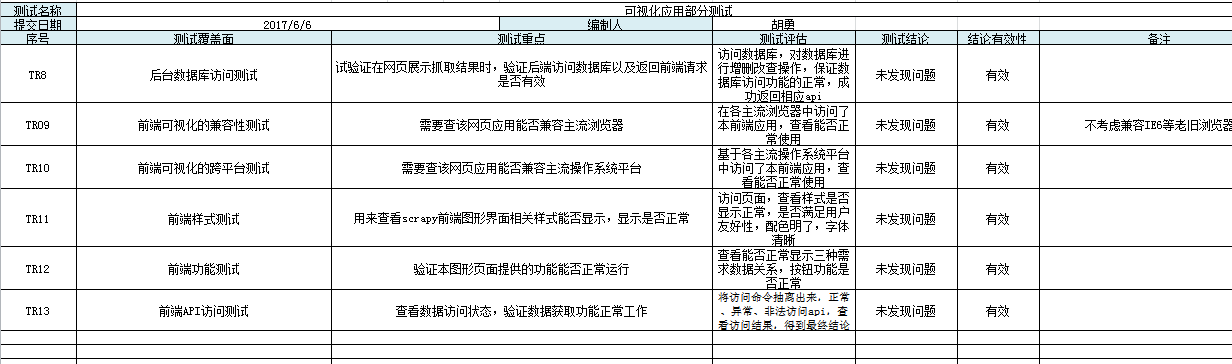
## 测试

* + 1. 覆盖范围

爬虫覆盖表如下截图：



可视化部分覆盖表如下截图：



* + 1. 代码规模

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 编号 | 子系统名称 | 构件名称 | 构建成员（语句/分支数量） | 说明章节（页号/行号） | 描述方式 | 简述/备注 |
| 1 | 过滤器 |  | 165行 | (A)1改进说明 | 文字说明 | 代码行数包含1.1,1.2 |
| 1.1 | 布隆过滤器 | 74行 | （A）1.2 （A）2.2.2 | 文字说明 | 涉及到filters.py |
| 1.2 | 分布式过滤器 | 91行 | （A）1.2  （A）2.2.1 | 文字说明 | 涉及到dupefilters.py |
| 2 | 调度器 | 调度器 | 365行 | （A）2.1.3 | 图简表 | 涉及到mongodb\_agent.py,  dupefilters.py,  scheduler.py |
| 3 | 网站画像 |  | 924行 | （A）2.1.4 | 图简表 | 代码行数包含3.1，3.2,3.3 |
| 3.1 | 爬虫部分 | 14行 | （A）2.1.4 | 图简表 | 该部分代码和调度器用到的mongodb代理是同一个程序文件。 |
| 3.2 | HTTP后端部分 | 370行 | （B）2.1  (B)2.2 | 文字说明，接口描述 |  |
| 3.3 | 前端展示部分 | 540行 | （C）1改进说明  （C） 技术说明  (C)展示说明 | 文字说明 |  |
| 4 | 反爬扩展 | 反扒扩展 | 120行 | （B）1.1  （B）1.2 | 文字说明 |  |
| 总计 |  |  | 1574行 |  |  |  |

## 1.3 结论

展现了分布式爬虫的爬取能力以及改进部分的正确性，通过测试覆盖表可以清楚的查看到改进与展现部分的测试已经全部覆盖，同时根据测试的结论可以大致了解改进部分的可用性。

# 进度和任务的计划与监控

## 方法

本小组的进度和任务的计划先由组长王益飞先进行下周进度的安排，然后由小组成员 各自上传本周工作日志或者跟组长提交本周工作日志，然后由组长天蝎完整本周具体完成内容。

## 效果

由于在实验的前几周时间，没有执行有效的计划和监控，导致前期的计划不能正确的表现出本组的工作量，但是，在之后有了具体的小组方法，小组进度和任务的计划能够很好的执行以及有效的监控整个实验的完成。