Proxy_Pool

Proxy_Pool，一个小巧的代理ip抓取+评估+存储+展示的一体化的工具，可自动化的搜集检测可用代理并进行评分，并添加了web展示和接口。

安装

1、从GitHub上脱下来，把代码放在web目录下。

```
git clone https://github.com/TideSec/Proxy_Pool
```

web服务器在unix/linux下可以用`https://github.com/teddysun/lamp`进行快速安装。

在windows下可以用phpstudy进行快速部署。

2、在mysql中新建数据库proxy，将proxy.sql文件导入，在include/config.inc.php中修改数据库密码。

3、此时本机访问http://ip:port，应该可以看到代理web展示界面

4、安装python2依赖库

```
pip install lxml
pip install requests
pip install pymysql
```

5、在py_proxy_task/config.py文件中配置数据库连接信息及其他参数。

使用

在py_proxy_task目录下有`proxy_get.py`和`proxy_check.py`两个程序，前者负责每天抓ip存进数据库，后者负责数据库中ip的清理和评估。

```
python proxy_get.py
# 等待上述程序抓取完结果后再运行评测程序
python proxy_check.py
```

之后按默认配置，这两个程序每天分别执行抓取和评估工作，放服务器上长期运行即可。

简介

原作者代码在这里：`https://github.com/chungminglu/Proxy`

我对部分代码进行了修改，完善了部分提取代理的解析代码，并加入了web展示和web接口，方便其他程序调用。

web页面我是从我的另外一个扫描器上改过来的`https://github.com/TideSec/WDScanner/`，里面可能有部分无用代码没有删除。

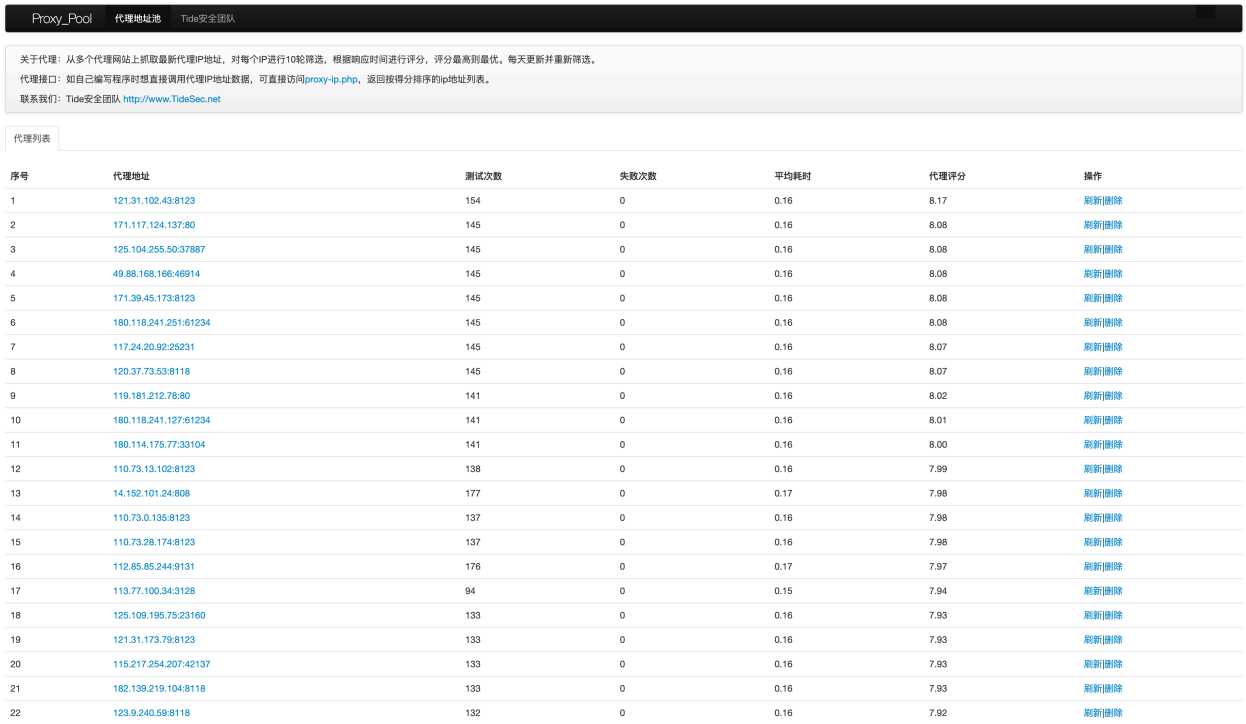
程序的几个功能：

1、每天从多个代理ip网站上抓下最新高匿ip数据。

2、经过筛选后的ip将存入数据库。

3、存入数据库的ip每天也要经过测试，存在剔除、评分机制，多次不合格的ip将被删除，每个ip都被评分，我们最终可以按得分排名获得稳定、低响应时间的优质ip。

web展示如下图所示：



Proxy_Pool 代理地址池 Tide安全团队

关于代理：从多个代理网站上抓取最新代理IP地址，对每个IP进行10轮筛选，根据响应时间进行评分，评分最高则最优。每天更新并重新筛选。

代理接口：如自己编写程序时想直接调用代理IP地址数据，可直接访问proxy-ip.php，返回按得分排序的ip地址列表。

联系我们：Tide安全团队 http://www.TideSec.net

代理列表

序号	代理地址	测试次数	失败次数	平均耗时	代理评分	操作
1	121.31.102.43:8123	154	0	0.16	8.17	刷新|删除
2	171.117.124.137:80	145	0	0.16	8.08	刷新|删除
3	125.104.255.50:37887	145	0	0.16	8.08	刷新|删除
4	49.88.168.166:46914	145	0	0.16	8.08	刷新|删除
5	171.39.45.173:8123	145	0	0.16	8.08	刷新|删除
6	180.118.241.251:61234	145	0	0.16	8.08	刷新|删除
7	117.24.20.92:25231	145	0	0.16	8.07	刷新|删除
8	120.37.73.53:8118	145	0	0.16	8.07	刷新|删除
9	119.181.212.78:80	141	0	0.16	8.02	刷新|删除
10	180.118.241.127:61234	141	0	0.16	8.01	刷新|删除
11	180.114.175.77:33104	141	0	0.16	8.00	刷新|删除
12	110.73.13.102:8123	138	0	0.16	7.99	刷新|删除
13	14.152.101.24:808	177	0	0.17	7.98	刷新|删除
14	110.73.0.135:8123	137	0	0.16	7.98	刷新|删除
15	110.73.28.174:8123	137	0	0.16	7.98	刷新|删除
16	112.85.85.244:9131	176	0	0.17	7.97	刷新|删除
17	113.77.100.34:3128	94	0	0.15	7.94	刷新|删除
18	125.109.195.75:23160	133	0	0.16	7.93	刷新|删除
19	121.31.173.79:8123	133	0	0.16	7.93	刷新|删除
20	115.217.254.207:42137	133	0	0.16	7.93	刷新|删除
21	182.139.219.104:8118	133	0	0.16	7.93	刷新|删除
22	123.9.240.59:8118	132	0	0.16	7.92	刷新|删除

web接口如下图所示：

www.123.com/proxy-ip.php

121.31.102.43:8123
171.117.124.137:80
125.104.255.50:37887
49.88.168.166:46914
171.39.45.173:8123
180.118.241.251:61234
117.24.20.92:25231
120.37.73.53:8118
119.181.212.78:80
180.118.241.127:61234
180.114.175.77:33104
110.73.13.102:8123

14.152.101.24:808
110.73.0.135:8123
110.73.28.174:8123
112.85.85.244:9131
113.77.100.34:3128
125.109.195.75:23160
121.31.173.79:8123
115.217.254.207:42137
182.139.219.104:8118
123.9.240.59:8118
113.69.128.37:8118
171.12.164.245:21967
121.12.42.158:61234
222.138.16.206:9999
121.232.146.220:9000
36.25.58.43:31299
110.73.40.12:8123
182.88.252.97:8123
42.177.0.145:9999
110.73.41.205:8123
110.72.40.118:8123
110.72.42.145:8123
110.73.29.184:8123
112.64.10.177:8118
115.225.254.166:808
110.72.26.39:8123
183.161.59.217:8998
115.203.213.21:808
110.73.42.127:8123
171.38.26.25:8123
110.73.6.149:8123
110.73.13.207:8123
125.112.172.75:33518
113.128.10.71:20128
110.73.54.222:8123
110.73.52.52:8123
1.196.2.49:33951

参数设置

在py_proxy_task/config.py文件可进行代理评估参数的设置。

```
USELESS_TIME = 4    # 最大失效次数
SUCCESS_RATE = 0.8
TIME_OUT_PENALTY = 10  # 超时惩罚时间
CHECK_TIME_INTERVAL = 24*3600  # 每天更新一次
```

除数据库配置参数外，主要用到的几个参数说明如下：

- `USELESS_TIME`和`SUCCESS_RATE`是配合使用的，当某个`ip`的`USELESS_TIME < 4 && SUCCESS_RATE < 0.8`时（同时兼顾到ip短期和长期的检测表现），则剔除该ip。
- `TIME_OUT_PENALTY`， 当某个ip在某次检测时失效，而又没有达到上一条的条件时（比如检测了100次后第一次出现超时），设置一个`response_time`的惩罚项，此处为10秒。
- `CHECK_TIME_INTERVAL`， 检测周期。此处设置为每隔12小时检测一次数据库里每一个ip的可用性。

策略

- 每天如下5个代理ip网站上抓下最新高匿ip数据：
  - `mimi`
  - `66ip`
  - `xici`
  - `cn-proxy`
  - `kuaidaili`
- N轮筛选
  - 收集到的ip集合将经过N轮，间隔为t的连接测试，对于每一个ip，必须全部通过这N轮测试才能最终进入数据库。如果当天进入数据库的ip较少，则暂停一段时间（一天）再抓。
- 数据库中ip评价准则
  - 检测过程中累计超时次数>`USELESS_TIME`&&成功率<`SUCCESS_RATE`就被剔除。
    `score = (success_rate + test_times / 500) / avg_response_time`
    原来的考虑是`score = success_rate / avg_response_time`，即：评分=成功率/平均响应时间， 考虑到检测合格过100次的老ip比新ip更有价值，检测次数也被引入评分。

关注我们

TideSec安全团队：

Tide安全团队正式成立于2019年1月，是以互联网攻防技术研究为目标的安全团队，目前聚集了十多位专业的安全攻防技术研究人员，专注于网络攻防、Web安全、移动终端、安全开发、IoT/物联网/工控安全等方向。

想了解更多Tide安全团队，请关注团队官网: http://www.TideSec.net 或关注公众号：