防爬虫案例

概述

当爬虫影响到我们网站的性能。

爬虫的种类：

1。善意的。百度，google。

2。恶意的。恶意窃取网站内容。

robots协议：

防爬虫的方法：现在爬虫ip，对我们系统的请求。

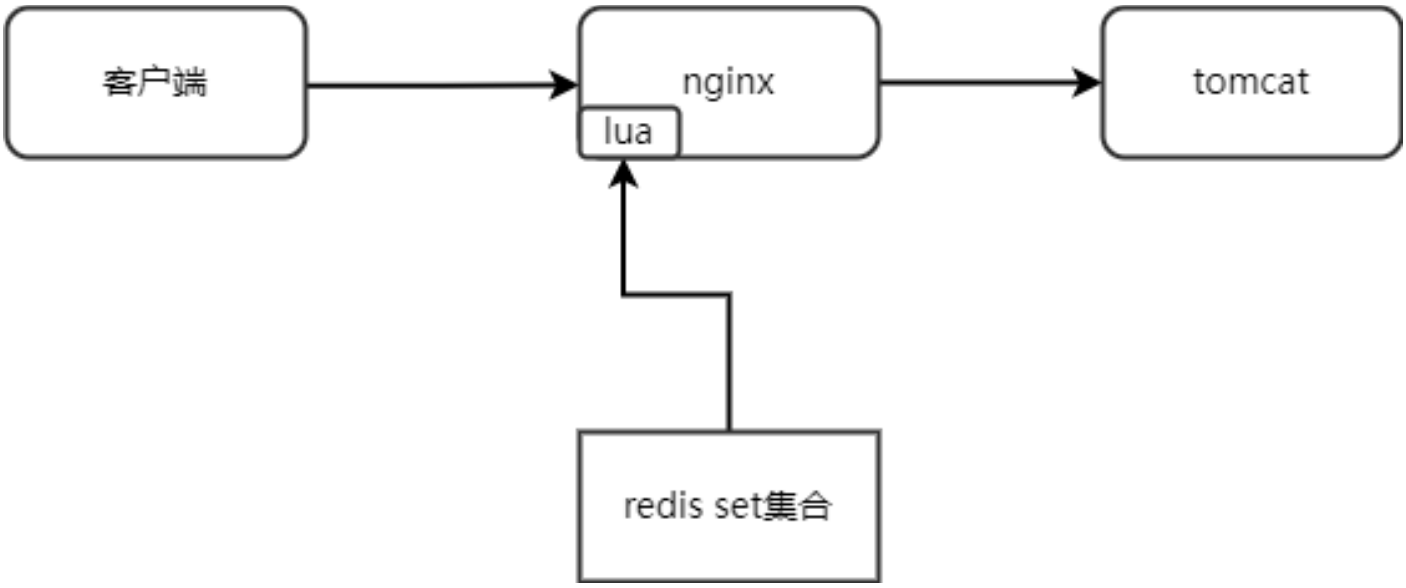
扩展：限制爬虫的方法：

1。限制user-agent。

2。限制ip。

3。添加验证码。

4。限制cookie



需求&步骤分解

1。收集黑名单IP。

2。存储到redis的set集合中。

3。nginx定期（2s）去从redis取 黑名单 ip 集合。

4。当请求来的时候，进行判断。请求来源的ip是否在ip黑名单中。

Redis黑名单准备

用set类型

key：ip-black-list

```
[root@localhost ~]# /usr/bin/redis-cli
127.0.0.1:6379> sadd ip-black-list 10.0.1.1
(integer) 1
127.0.0.1:6379> sadd ip-black-list 10.0.2.2
(integer) 1
127.0.0.1:6379> smember ip-balck-list
(error) ERR unknown command 'smember'
127.0.0.1:6379> smembers ip-balck-list
(empty list or set)
127.0.0.1:6379> smembers ip-black-list
1) "10.0.2.2"
2) "10.0.1.1"
127.0.0.1:6379>
```

编写nginx配置文件

```
[root@localhost nginx]# cat conf/nginx-black-list.conf
worker_processes 1;
error_log logs/error.log debug;

events {
    worker_connections 1024;
}

http {
    ## 定义共享空间
    lua_shared_dict ip_black_list 1m;
    include mime.types;
    default_type application/octet-stream;

    server {
        listen 8083;
        location / {
```

```
                default_type text/html;
                access_by_lua_file /usr/local/openresty/nginx/lua/black-list-access.lua;
                proxy_pass http://localhost:8080/;
            }
        }
}
[root@localhost nginx]#
```

共享变量：ip_black_list

编写核心lua脚本