带你了解代理 IP 那些事

极客猴

题图：by ourclickdays from Instagram

阅读文本大概需要 7 分钟。

在爬取某些网站时，我们经常会设置代理 IP 来避免爬虫程序被封。我们获取代理 IP 地址方式通常提取国内的知名 IP 代理商（如西刺代理，快代理，无忧代理等）的免费代理。这些代理商一般都会提供透明代理，匿名代理，高匿代理。那么这几种代理的区别是什么？我们该如何选择呢？本文的主要内容是讲解各种代理 IP 背后的原理。

1 代理类型

代理类型一共能分为四种。除了前面提到的透明代理，匿名代理，高匿代理，还有混淆代理。从安全程度来说，这四种代理类型的排序是 高匿 > 混淆 > 匿名 > 透明。

2 代理原理

代理类型主要取决于代理服务器端的配置。不同配置会形成不同的代理类型。在配置中，这三个变量 REMOTE_ADDR，HTTP_VIA，HTTP_X_FORWARDED_FOR 是决定性因素。

1）REMOTE_ADDR

REMOTE_ADDR 表示客户端的 IP，但是它的值不是由客户端提供的，而是服务器根据客户端的 IP 指定的。

如果使用浏览器直接访问某个网站，那么网站的 web 服务器（Nginx、Apache等）就会把 REMOTE_ADDR 设为客户端的 IP 地址。

如果我们给浏览器设置代理，我们访问目标网站的请求会先经过代理服务器，然后由代理服务器将请求转化到目标网站。那么网站的 web 服务器就会把 REMOTE_ADDR 设为代理服务器的 IP。

2）X-Forwarded-For（XFF）

X-Forwarded-For 是一个 HTTP 扩展头部，用来表示 HTTP 请求端真实 IP。当客户端使用了代理时，web 服务器就不知道客户端的真实 IP 地址。为了避免这个情况，代理服务器通常会增加一个 X-Forwarded-For 的头信息，把客户端的 IP 添加到头信息里面。

X-Forwarded-For 请求头格式如下：

```
X-Forwarded-For:  client,  proxy1,  proxy2
```

client 表示客户端的 IP 地址；proxy1 是离服务端最远的设备 IP；proxy2 是次级代理设备的 IP；从格式中，可以看出从 client 到 server 是可以有多层代理的。

如果一个 HTTP 请求到达服务器之前，经过了三个代理 Proxy1、Proxy2、Proxy3，IP 分别为 IP1、IP2、IP3，用户真实 IP 为 IP0，那么按照 XFF 标准，服务端最终会收到以下信息：

```
X-Forwarded-For:  IP0,  IP1,  IP2
```

Proxy3 直连服务器，它会给 XFF 追加 IP2，表示它是在帮 Proxy2 转发请求。列表中并没有 IP3，IP3 可以在服务端通过 Remote Address 字段获得。我们知道 HTTP 连接基于 TCP 连接，HTTP 协议中没有 IP 的概念，Remote Address 来自 TCP 连接，表示与服务端建立 TCP 连接的设备 IP，在这个例子里就是 IP3。

3）HTTP_VIA

via 是 HTTP 协议里面的一个header,记录了一次 HTTP 请求所经过的代理和网关，经过1个代理服务器，就添加一个代理服务器的信息，经过2个就添加2个。

3 代理类型区别

1）透明代理(Transparent Proxy)

代理服务器的配置如下：

```
REMOTE_ADDR = Proxy IP

HTTP_VIA = Proxy IP

HTTP_X_FORWARDED_FOR = Your IP
```

透明代理虽然可以直接“隐藏”客户端的 IP 地址，但是还是可以从HTTP_X_FORWARDED_FOR来查到客户端的 IP 地址。

2） 匿名代理(Anonymous Proxy)

代理服务器的配置如下：

```
REMOTE_ADDR = proxy IP

HTTP_VIA = proxy IP

HTTP_X_FORWARDED_FOR = proxy IP
```

匿名代理能提供隐藏客户端 IP 地址的功能。使用匿名代理，服务器能知道客户端使用用了代理，当无法知道客户端真实 IP 地址。

3） 混淆代理(Distorting Proxy)

代理服务器的配置如下：

```
REMOTE_ADDR = Proxy IP

HTTP_VIA = Proxy IP

HTTP_X_FORWARDED_FOR = Random IP address
```

与匿名代理的原理相似，但是会伪装得更逼真。如果客户端使用了混淆代理，服务器还是能知道客户端在使用代理，但是会得到一个假的客户端 IP 地址。

2） 高匿代理(Elite Proxy 或 High Anonymity Proxy)

代理服务器的配置如下：

```
REMOTE_ADDR = Proxy IP

HTTP_VIA = not determined

HTTP_X_FORWARDED_FOR = not determined
```

高匿代理既能让服务器不清楚客户端是否在使用代理，也能保证服务器获取不到客户端的真实 IP 地址。

4 代理的选择

普通匿名代理能隐藏客户机的真实 IP，但会改变我们的请求信息，服务器端有可能会认为我们使用了代理。不过使用此种代理时，虽然被访问的网站不能知道客户端的 IP 地址，但仍然可以知道你在使用代理，当然某些能够侦测 IP 的网页仍然可以查到客户端的 IP。

而高度匿名代理不改变客户机的请求，这样在服务器看来就像有个真正的客户浏览器在访问它，这时客户的真实IP是隐藏的，服务器端不会认为我们使用了代理。

因此，爬虫程序需要使用到代理 IP 时，尽量选择普通匿名代理和高匿名代理。另外，如果要保证数据不被代理服务器知道，推荐使用 HTTPS 协议的代理。

本文参考文章：

《HTTP 请求头中的 X-Forwarded-For》

http://gohom.win/2016/01/20/proxy-type/

《proxy代理类型:透明代理 匿名代理 混淆代理和高匿代理》

https://imququ.com/post/x-forwarded-for-header-in-http.html

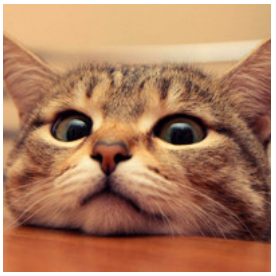
推荐阅读：

面对喜欢的人，该表达还是等待？

爬虫自学之路

人必有痴，而后有成

文章转载自公众号



极客猴

极客猴