**4.1.1微博移动版Web的分析**

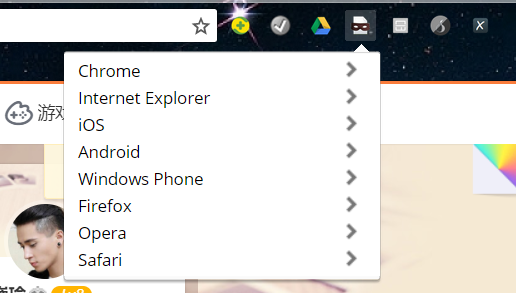
玩过爬虫的朋友应该都清楚，爬虫难度：www > m > wap （www是PC端，m和wap是移动端，现在的智能手机一般用的是m站，部分老手机用的还是wap），原因也很简单，现在的网站越来越多地使用AJAX加载，反爬虫机制也厉害。而像wap这种移动端网站限制比较小，网页结构也简单，我们获取、解析起来都简单很多，理论上速度也会快很多。所以如果允许的话我们尽量采用wap站抓取。



可能很多刚接触爬虫的朋友也想从wap爬取，但不知道怎么做。例如用PC端浏览器打开 http://weibo.com 在登录的时候会自动跳回m域名网站，甚至用requests打开网页时会返回403错误。   
这是因为网站服务器会根据你的浏览器表头判断你是从哪个平台发送的请求，识别到PC端的请求会给你作相应处理。所以我们只需要修改一下浏览器表头（User-Agent）即可。

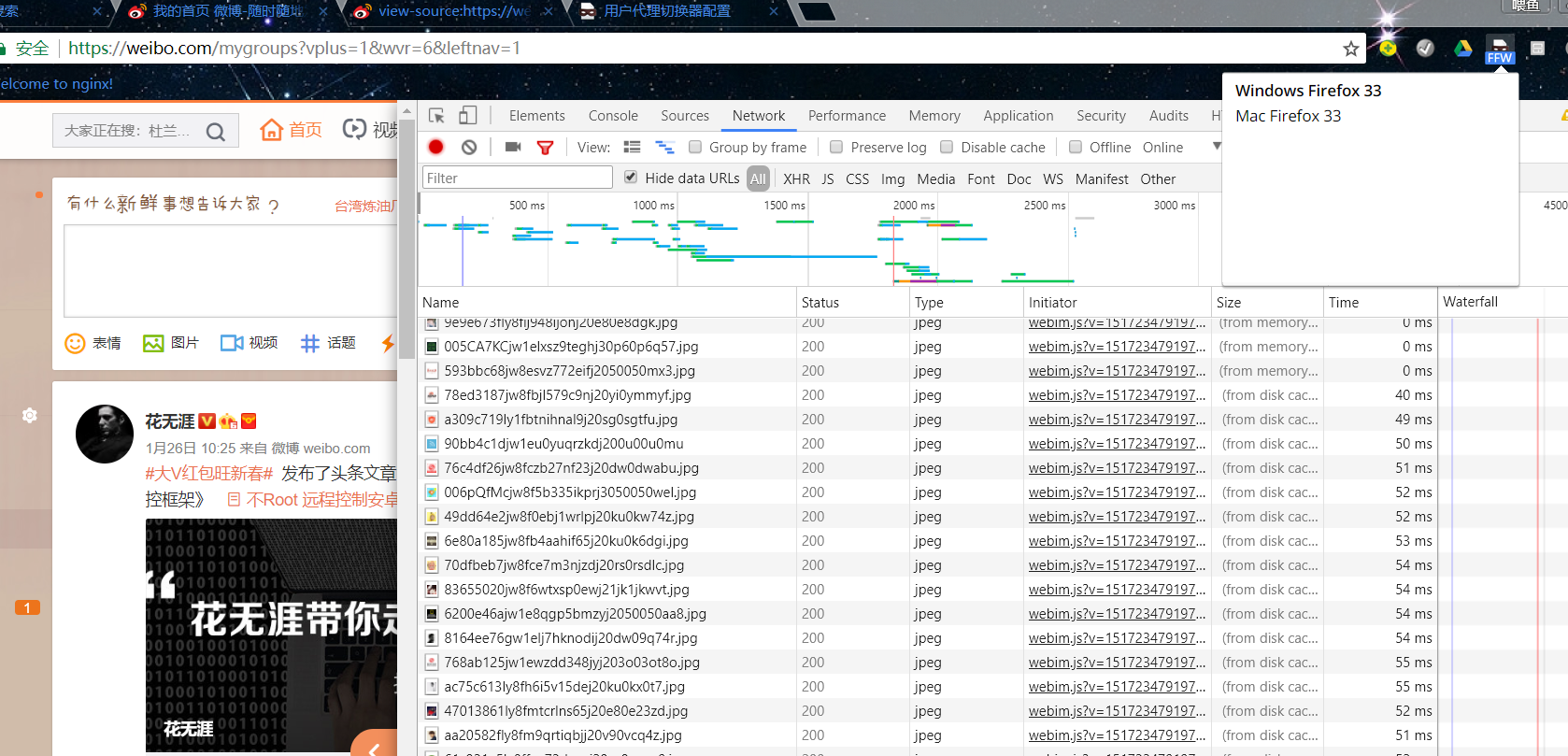
如果是爬虫程序，只需要带上旧版手机浏览器的User-Agent即可（例如：”Mozilla/5.0 (Linux; U; Android 2.3.6; en-us; Nexus S Build/GRK39F) AppleWebKit/533.1 (KHTML, like Gecko) Version/4.0 Mobile Safari/533.1”）。

然而我们只看程序返回的response内容并不舒爽，我们还想在PC端用浏览器模拟手机浏览器那样打开网页，怎么办？   
我们只需要把PC浏览器的User-Agent改成手机的User-Agent即可。   
例如我用的是Chrome50，修改浏览器的User-Agent有两种办法：一种是安装一个插件——User-agent Switcher，另一种是直接修改浏览器的表头（仅当前页面有效）。

User-agent Switcher插件：   


直接修改浏览器的User-Agent：

插件改为windows 火狐浏览器环境：



就新浪微博而言，打开一个微博用户的个人首页，wap站直接返回一个HTML文件，并不需要加载JS和CSS，而且格式、编码都很正常；而m站返回的内容格式比较混乱，用xpath解析不了（也有可能是我的程序有问题），而且使用的是Unicode编码格式。