**2.3.1网络爬虫引发的问题:**

* 网络爬虫在法律上的风险: 服务器上的数据有产权归属 网络爬虫获取数据后牟利将带来法律风险
* 网络爬虫在隐私的泄漏: 网络爬虫可能具备突破简单访问控制的能力，获得被保护数据 从而泄露个人隐私
* Web服务器默认接收人类访问 受限于编写水平和目的，网络爬虫将会为Web服务器带来巨大的资源开销

**2.3.2 Robots协议**：

Robots协议（也称为爬虫协议、机器人协议等）的全称是“网络爬虫排除标准”（Robots ExclusionProtocol），网站通过Robots协议告诉搜索引擎哪些页面可以抓取，哪些页面不能抓取.

作用：网站告知网络爬虫哪些页面可以抓取，哪些不行

形式：在网站根目录下的robots.txt文件

网络爬虫：自动或人工识别robots.txt，再进行内容爬取

约束性：Robots协议是建议但非约束性，网络爬虫可以不遵守，但存在法律风险

**1．）为什么需要Robots协议**

互联网上的网页是通过超级链接互相关联起来的，从而形成了网页的网状结构。爬虫的工作方式就像蜘蛛在网上沿着链接爬来爬去，最基本的流程可以简化如下：

1)喂给爬虫一堆URL，我们称之为种子(seeds)；

2)爬虫抓取seeds，解析html网页，抽取其中的超级链接；

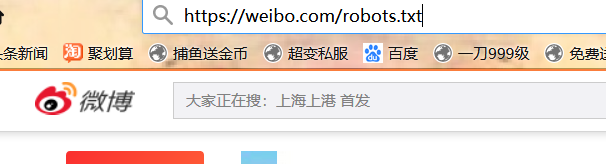
3)爬虫接着抓取这些新发现的链接指向的网页。

步骤2和步骤3循环往复。

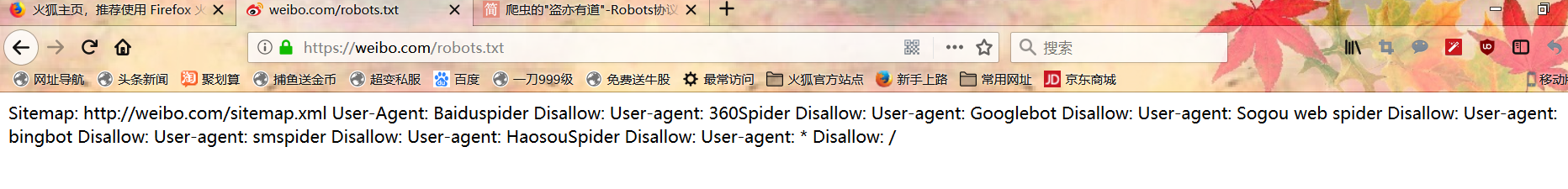
了解了上面的流程就能发现：对爬虫来说网站非常被动，只有老老实实被抓取的份。所以，对于网站的管理者来说，就存在这样的需求：

某些路径下是个人隐私或者网站管理使用，不想被搜索引擎抓取，比如说日本爱情动作片；不喜欢某个搜索引擎，不愿意被他抓取，最有名的就是之前淘宝不希望被百度抓取；小网站使用的是公用的虚拟主机，流量有限或者需要付费，希望搜索引擎抓的温柔点；某些网页是动态生成的，没有直接的链接指向，但是希望内容被搜索引擎抓取和索引。网站内容的所有者是网站管理员，搜索引擎应该尊重所有者的意愿，为了满足以上等等，就需要提供一种网站和爬虫进行沟通的途径，给网站管理员表达自己意愿的机会。有需求就有供应，robots协议就此诞生。

**2.）微博的Robots协议**

首先打开浏览器访问微博主页<https://weibo.com/> ,如下图:

在URL’/’后输入robots.txt回车 就能看到微博的robots协议如图所示:



就能看到微博的对爬虫的限制:

Sitemap: http://weibo.com/sitemap.xml User-Agent: Baiduspider Disallow: User-agent: 360Spider Disallow: User-agent: Googlebot Disallow: User-agent: Sogou web spider Disallow: User-agent: bingbot Disallow: User-agent: smspider Disallow: User-agent: HaosouSpider Disallow: User-agent: \* Disallow: /

显示360,bing,google,baidu,sogou,等等搜索引擎都不允许。