

爬虫简介

什么是爬虫？

通过编写程序，模拟浏览器上网，然后让其去互联网上

抓取数据的过程

爬虫的价值

----- | 实际应用 | 就业

合法性探究

不被法律禁止，但具有违法风险

爬虫初步深入

使用场景中的分类：

一 通用爬虫

抓取系统的重要组成部分，抓取的是一整张页面数据

二 聚集爬虫

建立在通用爬虫之上，抓取的是页面中特定的局部内容

三 增量式爬虫

监测网站中数据更新的情况，只会抓取网站中最新

更新出来的数据

爬虫的矛与盾

... ...

反爬机制

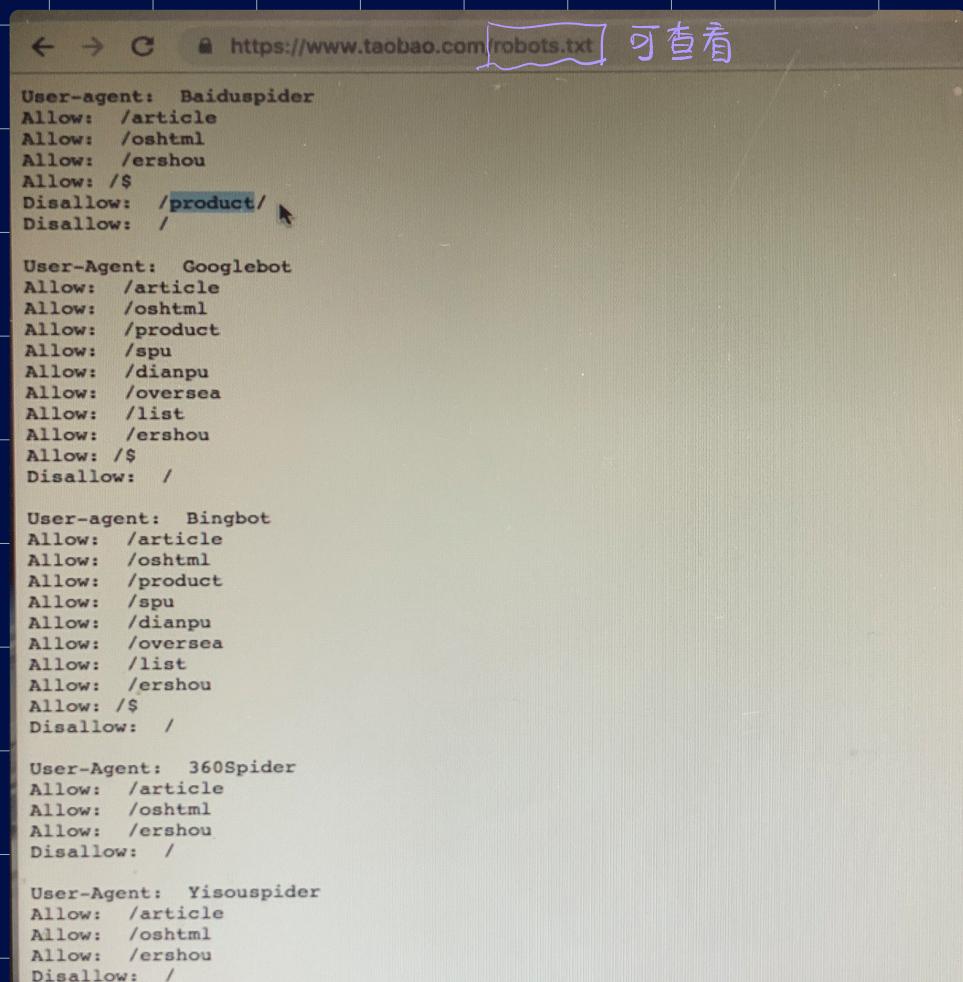
门户网站，通过制定相对应的策略或技术手段，防止程序抓取

反爬机制

~~~

### robots.txt 协议

“君子”协议。明确规定哪些内容可或不可被爬取。



A screenshot of a web browser displaying the robots.txt file for the website <https://www.taobao.com>. The page shows the following content:

```
User-agent: Baiduspider
Allow: /article
Allow: /oshtml
Allow: /ershou
Allow: /$
Disallow: /product/
Disallow: /

User-Agent: Googlebot
Allow: /article
Allow: /oshtml
Allow: /product
Allow: /spu
Allow: /dianpu
Allow: /oversea
Allow: /list
Allow: /ershou
Allow: /$
Disallow: /

User-agent: Bingbot
Allow: /article
Allow: /oshtml
Allow: /product
Allow: /spu
Allow: /dianpu
Allow: /oversea
Allow: /list
Allow: /ershou
Allow: /$
Disallow: /

User-Agent: 360Spider
Allow: /article
Allow: /oshtml
Allow: /ershou
Disallow: /

User-Agent: Yisouspider
Allow: /article
Allow: /oshtml
Allow: /ershou
Disallow: /
```

The word "查看" (View) is handwritten next to the URL in the address bar.

http & https 协议

http 协议

- 根概念：服务器与客户端进行数据交互的一种形式

常用请求头信息

- User-Agent：请求载体的身份标识

- Connection：请求完毕后，是断开连接还是保持

常用响应头信息

- Content-Type：服务器响应给客户端的数据类型

https 协议

“s” 代表 security，安全的 http (超文本传输协议)

数据加密！

加密方式

- 对称加密
- 非对称加密
- 许书本加密 ← https 所采用