<http://www.jianshu.com/p/73697828b8f8>

那么爬虫进阶应该如何做呢？通常所谓的进阶有以下几种：

**分布式**

通常会有一些教材告诉你，为了爬取效率，需要把爬虫分布式部署到多台机器上。这完全是骗人的。分布式唯一的作用是：防止对方封IP。封IP是终极手段，效果非常好，当然，误伤起用户也是非常爽的。

**模拟JavaScript**

有些教程会说，模拟javascript，抓取动态网页，是进阶技巧。但是其实这只是个很简单的功能。因为，如果对方没有反爬虫，你完全可以直接抓ajax本身，而无需关心js怎么处理的。如果对方有反爬虫，那么javascript必然十分复杂，重点在于分析，而不仅仅是简单的模拟。

换句话说：这应该是基本功。

**PhantomJs**

这个是一个极端的例子。这个东西本意是用来做自动测试的，结果因为效果很好，很多人拿来做爬虫。但是这个东西有个硬伤，就是：效率。此外PhantomJs也是可以被抓到的，出于多方面原因，这里暂时不讲。