Atitit 爬虫的问题与疑问qa

爬虫的功能来说。用户比较关心的问题往往是：  
　　1）爬虫支持多线程么、爬虫能用代理么、爬虫会爬取重复数据么、爬虫能爬取JS生成的信息么？  
　　不支持多线程、不支持代理、不能过滤重复URL的，那都不叫开源爬虫，那叫循环执行http请求。  
　　能不能爬js生成的信息和爬虫本身没有太大关系。爬虫主要是负责遍历网站和下载页面。爬js生成的信息和网页信息抽取模块有关，往往需要通过模拟浏览器（htmlunit,selenium）来完成。这些模拟浏览器，往往需要耗费很多的时间来处理一个页面。所以一种策略就是，使用这些爬虫来遍历网站，遇到需要解析的页面，就将网页的相关信息提交给模拟浏览器，来完成JS生成信息的抽取。  
　　2）爬虫可以爬取ajax信息么？  
　　网页上有一些异步加载的数据，爬取这些数据有两种方法：使用模拟浏览器（问题1中描述过了），或者分析ajax的http请求，自己生成ajax请求的url，获取返回的数据。如果是自己生成ajax请求，使用开源爬虫的意义在哪里？其实是要用开源爬虫的线程池和URL管理功能（比如断点爬取）。  
　　如果我已经可以生成我所需要的ajax请求（列表），如何用这些爬虫来对这些请求进行爬取？  
　　爬虫往往都是设计成广度遍历或者深度遍历的模式，去遍历静态或者动态页面。爬取ajax信息属于deep web（深网）的范畴，虽然大多数爬虫都不直接支持。但是也可以通过一些方法来完成。比如WebCollector使用广度遍历来遍历网站。爬虫的第一轮爬取就是爬取种子集合（seeds）中的所有url。简单来说，就是将生成的ajax请求作为种子，放入爬虫。用爬虫对这些种子，进行深度为1的广度遍历（默认就是广度遍历）。  
　　3）爬虫怎么爬取要登陆的网站？  
　　这些开源爬虫都支持在爬取时指定cookies，模拟登陆主要是靠cookies。至于cookies怎么获取，不是爬虫管的事情。你可以手动获取、用http请求模拟登陆或者用模拟浏览器自动登陆获取cookie。  
　　4）爬虫怎么抽取网页的信息？  
　　开源爬虫一般都会集成网页抽取工具。主要支持两种规范：CSS SELECTOR和XPATH。至于哪个好，这里不评价。  
　　5）爬虫怎么保存网页的信息？  
　　有一些爬虫，自带一个模块负责持久化。比如webmagic，有一个模块叫pipeline。通过简单地配置，可以将爬虫抽取到的信息，持久化到文件、数据库等。还有一些爬虫，并没有直接给用户提供数据持久化的模块。比如crawler4j和webcollector。让用户自己在网页处理模块中添加提交数据库的操作。至于使用pipeline这种模块好不好，就和操作数据库使用ORM好不好这个问题类似，取决于你的业务。  
　　6）爬虫被网站封了怎么办？  
　　爬虫被网站封了，一般用多代理（随机代理）就可以解决。但是这些开源爬虫一般没有直接支持随机代理的切换。所以用户往往都需要自己将获取的代理，放到一个全局数组中，自己写一个代理随机获取（从数组中）的代码。  
　　7）网页可以调用爬虫么？  
　　爬虫的调用是在Web的服务端调用的，平时怎么用就怎么用，这些爬虫都可以使用。  
　　8）爬虫速度怎么样？  
　　单机开源爬虫的速度，基本都可以讲本机的网速用到极限。爬虫的速度慢，往往是因为用户把线程数开少了、网速慢，或者在数据持久化时，和数据库的交互速度慢。而这些东西，往往都是用户的机器和二次开发的代码决定的。这些开源爬虫的速度，都很可以。  
　　9）明明代码写对了，爬不到数据，是不是爬虫有问题，换个爬虫能解决么？  
　　如果代码写对了，又爬不到数据，换其他爬虫也是一样爬不到。遇到这种情况，要么是网站把你封了，要么是你爬的数据是javascript生成的。爬不到数据通过换爬虫是不能解决的。  
　　10）哪个爬虫可以判断网站是否爬完、那个爬虫可以根据主题进行爬取？  
　　爬虫无法判断网站是否爬完，只能尽可能覆盖。  
　　至于根据主题爬取，爬虫之后把内容爬下来才知道是什么主题。所以一般都是整个爬下来，然后再去筛选内容。如果嫌爬的太泛，可以通过限制URL正则等方式，来缩小一下范围。