* **通过UA 识别爬虫**

有些爬虫的UA是特殊的，与正常浏览器的不一样，可通过识别特征UA，直接封掉爬虫请求

* **设置IP访问频率，如果超过一定频率，弹出验证码**

如果输入正确的验证码，则放行，如果没有输入，则拉入禁止一段时间，如果超过禁爬时间，再次出发验证码，则拉入黑名单。当然根据具体的业务，为不同场景设置不同阈值，比如登陆用户和非登陆用户，请求是否含有refer。

* **通过并发识别爬虫**

有些爬虫的并发是很高的，统计并发最高的IP，加入黑名单（或者直接封掉爬虫IP所在C段）

* **请求的时间窗口过滤统计**

爬虫爬取网页的频率都是比较固定的，不像人去访问网页，中间的间隔时间比较无规则，所以我们可以给每个IP地址建立一个时间窗口，记录IP地址最近12次访问时间，每记录一次就滑动一次窗口，比较最近访问时间和当前时间，如果间隔时间很长判断不是爬虫，清除时间窗口，如果间隔不长，就回溯计算指定时间段的访问频率，如果访问频率超过阀值，就转向验证码页面让用户填写验证码

* **限制单个ip/api token的访问量**

比如10分钟限制访问页面100次，具体标准可参考一些大型网站的公开api，对于抓取用户公开信息的爬虫要格外敏感

* **识别出合法爬虫**

对http头agent进行验证，是否标记为、百度的spider，严格一点的话应该判别来源IP是否为、baidu的爬虫IP，这些IP在网上都可以找到。校验出来IP不在白名单就可以阻止访问内容。

* **蜜罐资源**

爬虫解析离不开正则匹配，适当在页面添加一些正常浏览器浏览访问不到的资源，一旦有ip访问，过滤下头部是不是搜素引擎的蜘蛛，不是就可以直接封了。比如说隐式链接。