- 模式检测

  文件系统的符号连接和类似的错误配置所造成的环路会遵循某种模式；比如，URL 会随着组件的复制逐渐增加。有些机器人会将具有重复组件的 URL 当作潜在的环路，拒绝爬行带有多于两或三个重复组件的 URL。

  重复并不都是立即出现的（比如，“/subdir/subdir/subdir...”）。有些环路周期可能为 2 或其他间隔，比如“/subdir/images/subdir/images/subdir/images/...”。有些机器人会查找具有几种不同周期的重复模式。

- 内容指纹

  一些更复杂的 Web 爬虫会使用指纹这种更直接的方式来检测重复。使用内容指纹的机器人会获取页面内容中的字节，并计算出一个校验和（checksum）。这个校验和是页面内容的压缩表示形式。如果机器人获取了一个页面，而此页面的校验和它曾经见过，它就不会再去爬行这个页面的链接了——如果机器人以前见过页面的内容，它就已经爬行过页面上的链接了。

  必须对校验和函数进行选择，以求两个不同页面拥有相同校验和的几率非常低。MD5 这样的报文摘要函数就常被用于指纹计算。

  有些 Web 服务器会在传输过程中对页面进行动态的修改，所以有时机器人会在校验和的计算中忽略 Web 页面内容中的某些部分，比如那些嵌入的链接。而且，无论定制了什么页面内容的动态服务器端包含（比如添加日期、访问计数等）都可能会阻碍重复检测。

- 人工监视

  Web 就是一片荒野。勇敢的机器人最终总会陷入一个采用任何技术都无能为力的困境。设计所有产品级机器人时都要有诊断和日志功能，这样人类才能很方便地监视机器人的进展，如果发生了什么不寻常的事情就可以很快收到警告。在某些情况下，愤怒的网民会给你发送一些无礼的邮件来提示你出了问题。

爬行 Web 这样规模庞大的数据集时，好的蜘蛛探测法总是会不断改进其工作的。随着新的资源类型不断加入 Web，它会随着时间的推移构建出一些新的规则，并采纳这些规则。好的规则总是在不断发展之中的。

当受到错误爬虫影响的资源（服务器、网络带宽等）处于可管理状态，或者处于执行爬行工作的人的控制之下（比如在内部站点上）时，很多较小的、更具个性的爬虫就会绕开这些问题。这些爬虫更多的是依赖人类的监视来防止这些问题的发生。

224