- 规范化 URL
  将 URL 转换为标准形式以避免语法上的别名

- 广度优先的爬行
  每次爬虫都有大量潜在的 URL 要去爬行。以广度优先的方式来调度 URL 去访问 Web 站点，就可以将环路的影响最小化。即使碰到了机器人陷阱，也可以在回到环路中获取的下一个页面之前，从其他 Web 站点中获取成百上千的页面。如果采用深度优先方式，一头扎到单个站点中去，就可能会跳入环路，永远无法访问其他站点。[6]

- 节流[7]
  限制一段时间内机器人可以从一个 Web 站点获取的页面数量。如果机器人跳进了一个环路，试图不断地访问某个站点的别名，也可以通过节流来限制重复的页面总数和对服务器的访问总数。

- 限制 URL 的大小
  机器人可能会拒绝爬行超出特定长度（通常是 1KB）的 URL。如果环路使 URL 的长度增加，长度限制就会最终终止这个环路。有些 Web 服务器在使用长 URL 时会失败，因此，被 URL 增长环路困住的机器人会使某些 Web 服务器崩溃。这会让网管错误地将机器人当成发起拒绝服务攻击的攻击者。

  要小心，这种技术肯定会让你错过一些内容。现在很多站点都会用 URL 来管理用户的状态（比如，在一个页面引用的 URL 中存储用户 ID）。用 URL 长度来限制爬虫可能会带来些麻烦；但如果每当请求的 URL 达到某个特定长度时，都记录一次错误的话，就可以为用户提供一种检查某特定站点上所发生情况的方法。

- URL/ 站点黑名单
  维护一个与机器人环路和陷阱相对应的已知站点及 URL 列表，然后像躲避瘟疫一样避开它们。发现新问题时，就将其加入黑名单。

  这就要求有人工进行干预。但现在很多大型爬虫产品都有某种形式的黑名单，用于避开某些存在固有问题或者有恶意的站点。还可以用黑名单来避开那些对爬行大惊小怪的站点。[8]

223

注 6：总之，广度优先搜索是个好方法，这样可以更均匀地分配请求，而不是都压到任意一台服务器上去。这样可以帮助机器人将用于一台服务器的资源保持在最低水平。

注 7：在 9.5 节也讨论了请求率的节流问题。

注 8：9.4 节讨论了站点怎样才能避免被爬行，但有些用户拒绝使用这种简单的控制机制，在其站点被爬行时又会变得非常愤怒。