（续）

操作指南	描　述
制约自己	机器人应该记录访问每个站点的次数以及访问的时间，并通过这些信息来确保它没有太频繁地访问某个站点。机器人访问站点的频率高于几分钟一次时，管理员就要起疑心了。机器人每隔几秒钟就访问一次站点时，有些管理员就会生气了。机器人尽可能频繁地去访问一个站点，将所有其他流量都拒之门外时，管理员就会暴怒起来。 总之，应该限制机器人，使其每分钟最多只发送几条请求，并确保每条请求之间有几秒钟的间隔。还应该限制对站点的访问总次数，以防止环路的出现
(4) 容忍存在环路、重复和其他问题	
处理所有返回代码	必须做好处理所有 HTTP 状态码的准备，包括各种重定向和错误码。还应该对这些代码进行记录和监视。如果某站点出现大量不成功的结果，就应该对其进行调查。可能是很多 URL 过期了，或者服务器拒绝向机器人提供这些文档
规范 URL	试着将所有 URL 都转化为标准形式来消除常见的别名
积极地避免环路的出现	努力地检测并避免环路的出现。将操纵爬虫的过程当作一个反馈回路。应该将问题的结果和解决方法回馈到下一次爬行中，使爬虫在每次迭代之后都能表现得更好
监视陷阱	有些环路是故意造成的恶意环路。这些环路可能很难检测。有的站点会带有一些怪异的 URL，要监视对这类站点进行的大量访问。这种情况可能就是陷阱
维护一个黑名单	找到陷阱、环路、故障站点和不希望机器人访问的站点时，要将其加入一个黑名单，不要再次访问这些站点
(5) 可扩展性	
了解所需空间	事先通过数学计算明确你要解决的问题规模有多大。你可能会对应用程序完成一项机器人任务所需的内存规模感到非常吃惊，这是由 Web 庞大的规模造成的
了解所需带宽	了解你有多少网络带宽可用，以及在要求的时间内完成机器人任务所需的带宽大小。监视网络带宽的实际使用情况。你很可能会发现输出带宽（请求）要比输入带宽（响应）小得多。通过对网络使用情况的监视，可能还会找到一些方法来更好地优化你的机器人，通过更好地使用其 TCP 连接更好地利用网络带宽 [26]
了解所需的时间	了解机器人完成其任务所需花费的时间，检查这个进度是否与自己的估计相符。如果机器人的耗费与自己的估计相去甚远，可能就会有问题，需要进行调查

注 26：更多有关 TCP 性能优化的内容请参见第 4 章。