（续）

操作指南	描　述
分而治之	对大规模的爬行来说，很可能需要使用更多的硬件来完成这项工作，可以使用带有多个网卡的大型多处理器服务器，也可以使用多台较小的计算机共同配合工作
(6) 可靠性	
彻底测试	在将机器人放出去之前，要对其进行彻底的内部测试。作好非现场测试准备时，要先进行几次小型的处女航。收集大量结果并对性能和内存使用情况进行分析，估计一下它们会怎样累积成较大问题
检查点	所有严谨的机器人都要保存其进展的快照，出现故障时可以从那里重新开始。故障总是存在的：你会发现一些软件的 bug，硬件也会出故障。大规模机器人不能在每次出现这种情况时都从头开始。一开始就要设计检查点/重启机制
故障恢复	预测故障的发生，对机器人进行设计，使其能够在发生故障时继续工作
(7) 公共关系	
做好准备	机器人可能会让很多人感到困惑。要作好快速响应其询问的准备。制定一个 Web 页面政策声明，对机器人进行描述，其中包括创建 robots.txt 文件的详细指南
充分理解	有些与你联系，讨论机器人问题的人是了解情况并赞成的，有些人则很幼稚。少数人会异常愤怒。有些人看起来好像都要发疯了。去争辩机器人的努力有多么重要通常是没什么效果的。向他们解释拒绝机器人访问标准，如果他们仍然很不高兴，就立即将投诉者的 URL 从爬行列表中删除，并将其加入黑名单
积极响应	大多数不满意的网管都只是不太了解机器人。如果你能够进行迅速且专业的响应，90% 的投诉都会很快消失。另一方面，如果你等好几天才响应，而机器人在继续访问这个站点，你面对的就将是一个非常愤怒的对手

241

9.6 搜索引擎

得到最广泛使用的 Web 机器人都是因特网搜索引擎。因特网搜索引擎可以帮助用户找到世界范围内涉及任意主题的文档。

现在 Web 上很多最流行的站点都是搜索引擎。很多 Web 用户将其作为起始点，它们会为用户提供宝贵的服务，帮助用户找到他们感兴趣的信息。

Web 爬虫为因特网搜索引擎提供信息，它们获取 Web 上的文档，并允许搜索引擎创建与本书后面的索引类似的索引，用以说明哪些文档中有哪些词存在。搜索引擎是 Web 机器人的主要来源——让我们来快速了解一下它们是如何工作的。