表9-6 Web机器人操作员指南

操作指南	描　述
(1) 识别	
识别你的机器人	用 HTTP 的 User-Agent 字段将机器人的名字告诉 Web 服务器。这样可以帮助管理员理解机器人所做的事情。有些机器人还会在 User-Agent 首部包含一个描述机器人目的和策略的 URL
识别你的机器	确保机器人是从一台带有 DNS 条目的机器上运行的，这样 Web 站点才能够将机器人的 IP 地址反向 DNS 为主机名。这有助于管理者识别出对机器人负责的组织
239 识别联络人	用 HTTP 的 From 字段提供一个联络的 E-mail 地址
(2) 操作	
保持警惕	机器人可能会惹一些麻烦或引发一些抱怨。其中一些是由那些行为有偏差的机器人造成的。一定要小心，注意保持机器人的正常行为。如果机器人要全天候运行，就要格外小心。需要有操作人员不间断地对机器人进行监视，直到它有了丰富的经验为止
做好准备	开始一次重要的机器人之旅时，一定要通知你所在的组织。你的组织可能要观测网络带宽的耗费，作好应对各种公共查询的准备
监视并记录日志	机器人应该装备有丰富的诊断和日志记录工具，这样才能记录进展、识别所有的机器人陷阱，进行完整性检查看看工作是否正常。监视并记录机器人行为的重要性怎么强调也不过分。问题和抱怨总是会有的，对爬虫行为的详细记录，有助于机器人操作者回溯所发生的事情。不管是为了调试出错的 Web 爬虫，还是为了在不合理的投诉面前为其行为进行辩护，监视和记录工作都是非常重要的
学习并适应	在每次爬行中你都会学到新的东西。要让机器人逐步适应，这样，它在每次爬行之后都会有所进步，并能避开一些常见的陷阱
(3) 约束自己的行为	
对 URL 进行过滤	如果一个 URL 指向的好像是你不理解或不感兴趣的数据，你可能会希望跳过它。比如，以 .Z、.gz、.tar 或者 .zip 结尾的 URL 很可能是压缩文件或归档文件。以 .exe 结尾的 URL 可能就是程序。以 .gif、.tif、.jpg 结尾的 URL 很可能是图片。要确保你得到的就是你想要的
过滤动态 URL	通常，机器人不会想去爬行来自动态网关的内容。机器人不知道应该如何正确地格式化查询请求，并将其发送给网关，而它得到的结果也很可能是错误的或临时的。如果一个 URL 中包含了 cgi，或者有一个“?”，机器人可能就不会去爬行这个 URL 了
对 Accept 首部进行过滤	机器人应该用 HTTP 的 Accept 首部来告诉服务器它能够理解哪种内容
遵循 robots.txt	机器人应该接受站点上 robots.txt 的控制