管理员提供了一种能够更好地控制机器人行为的机制。这个标准被称为“拒绝机器人访问标准”，但通常只是根据存储访问控制信息的文件而将其称为 robots.txt。

robots.txt 的思想很简单。所有 Web 服务器都可以在服务器的文档根目录中提供一个可选的、名为 robots.txt 的文件。这个文件包含的信息说明了机器人可以访问服务器的哪些部分。如果机器人遵循这个自愿约束标准，它会在访问那个站点的所有其他资源之前，从 Web 站点请求 robots.txt 文件。例如，图 9-6 中的机器人想要从 Joe 的五金商店下载 http://www.joes-hardware.com/specials/acetylene-torches.html。但在机器人去请求这个页面之前，要先去查看 robots.txt 文件，看看它是否有获取这个页面的权限。在这个例子中，robots.txt 文件并没有拦截机器人，因此机器人获取了这个页面。

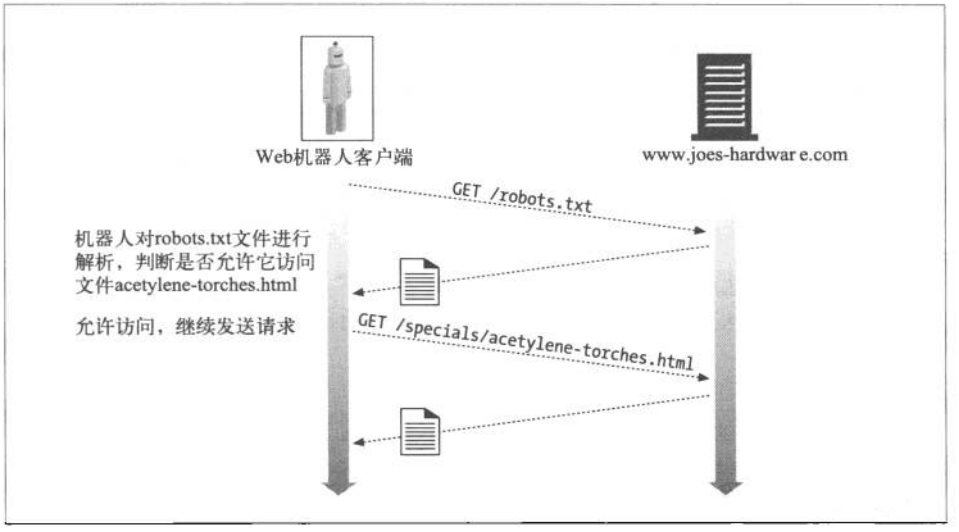

图 9-6　在爬行目标文件之前，先获取 robots.txt，验证是否可以进行访问

9.4.1　拒绝机器人访问标准

拒绝机器人访问标准是一个临时标准。编写本书的时候还没有官方标准机构承认这个标准，不同的厂商实现了这个标准的不同子集。但是，具备一些对机器人访问 Web 站点的管理能力，即使并不完美，也总比一点儿都没有要好，而且大部分主要的生产厂商和搜索引擎爬虫都支持这个拒绝访问标准。

尽管没有很好地定义版本的名称，但拒绝机器人访问标准是有三个版本的。我们采用了表 9-2 列出的版本编号。