经过这些步骤就可以消除表 9-1 中 a~c 所列的别名问题了。但如果不知道特定 Web 服务器的相关信息，机器人就没什么好办法来避免表 9-1 中 d~f 的问题了。

- 机器人需要知道 Web 服务器是否是大小写无关的才能避免表 9-1d 中的别名问题。
- 机器人需要知道 Web 服务器上这个目录下的索引页面配置才能知道表 9-1e 中的情况是否是别名。
- 即使机器人知道表 9-1f 中的主机名和 IP 地址都指向同一台计算机，它也还要知道 Web 服务器是否配置为进行（参见第 5 章）虚拟主机操作，才能知道这个 URL 是不是别名。

URL 规范化可以消除一些基本的语法别名，但机器人还会遇到其他的、将 URL 转换为标准形式也无法消除的 URL 别名。

9.1.8 文件系统连接环路

文件系统中的符号连接会造成特定的潜在环路，因为它们会在目录层次深度有限的情况下，造成深度无限的假象。符号连接环路通常是由服务器管理员的无心错误造成的，但“邪恶的网管”也可能会恶意地为机器人制造这样的陷阱。

图 9-3 显示了两个文件系统。在图 9-3a 中，subdir 是个普通的目录。在图 9-3b 中，subdir 是个指回到“/”的符号连接。在这两个图中，都假设文件 /index.html 中包含了一个指向文件 subdir/index.html 的超链。

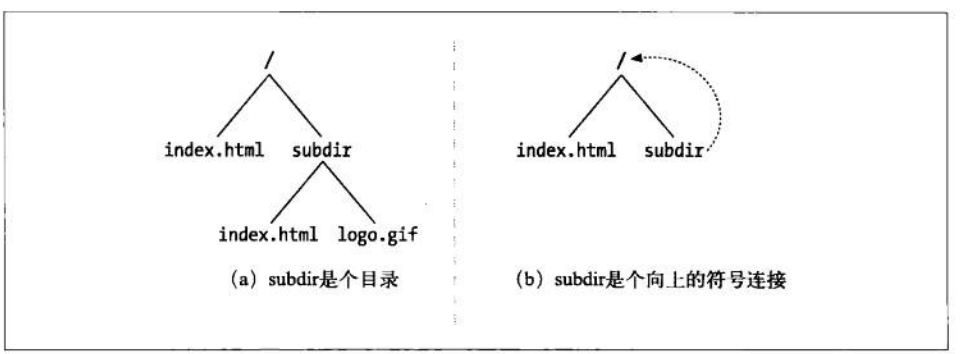

图 9-3 符号连接环路

使用图 9-3a 所示的文件系统时，Web 爬虫可能会采取下列动作：

(1) GET http://www.foo.com/index.html

获取 /index.html，找到指向 subdir/index.html 的链接。