包含 Host 首部的话，可能会使机器人将错误的内容与一个特定的 URL 关联起来。因此，HTTP/1.1 要求使用 Host 首部。

在默认情况下，大多数服务器都被配置为提供一个特定的站点。因此，不包含 Host 首部的爬虫向提供两个站点的服务器发起请求时，就像图 9-5 中的站点一样（www.joes-hardware.com 和 www.foo.com），假设默认情况下服务器被配置为提供 www.joes-hardware.com 站点（且不需要 Host 首部），那么，若请求 www.foo.com 上的某个页面，爬虫实际获取的就是 Joe 的五金商店的站点上的内容。更糟糕的是，爬虫会认为来自 Joe 的五金站点上的那些内容是来自 www.foo.com 的。如果带有相对立的政治色彩或其他观点的两个站点是由同一台服务器提供的，你肯定能想象到会有更不幸的局面出现。

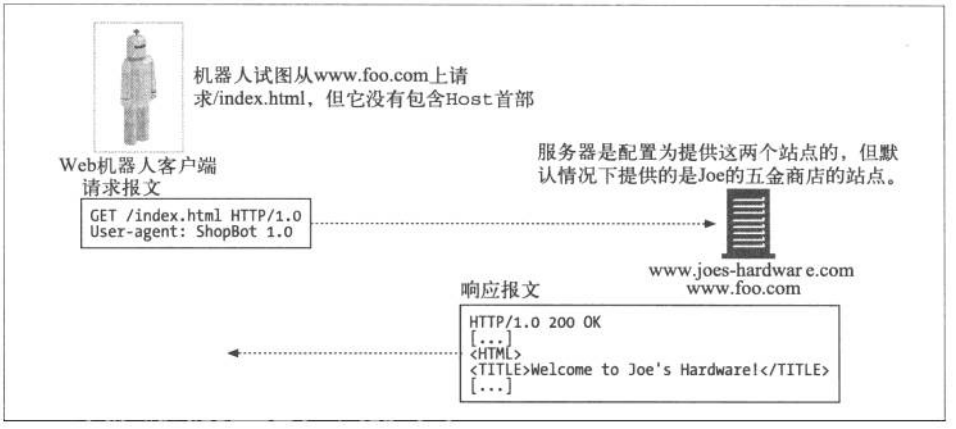

图 9-5　发送请求时没有携带 Host 首部，虚拟 docroot 会引发问题的例子

9.2.3　条件请求

鉴于这些机器人的努力程度，尽量减少机器人所要获取内容的数量通常是很有意义的。对因特网搜索引擎机器人来说，需要下载的潜在页面有数十亿，所以，只在内容发生变化时才重新获取内容是很有意义的。

有些机器人实现了条件 HTTP 请求，[12] 它们会对时间戳或实体标签进行比较，看看它们最近获取的版本是否已经升级了。这与 HTTP 缓存查看已获取资源的本地副本是否有效的方法非常相似。更多与缓存对资源本地副本的验证有关的信息请参见第 7 章。

226

注 12：3.5.2 节给出了一个机器人可以实现的条件首部的完整列表。