一个指向下个月的链接。真正的用户是不会不停地请求下个月的链接的，但不了解其内容的动态特性的机器人可能会不断向这些资源发出无穷的请求。[5]

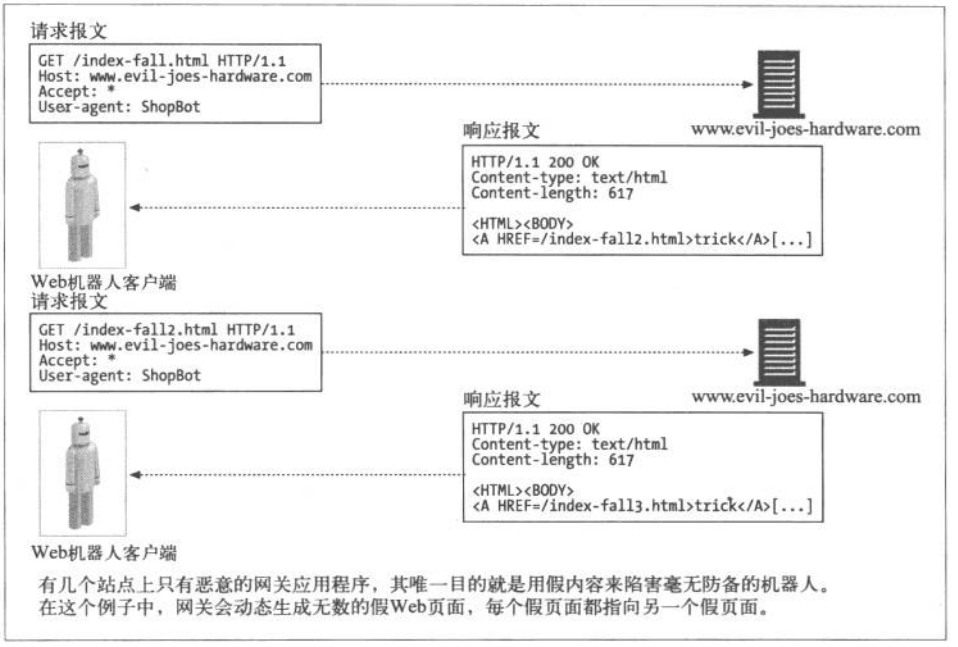

图 9-4　恶意的动态 Web 空间示例

9.1.10　避免循环和重复

没有什么简单明了的方式可以避免所有的环路。实际上，经过良好设计的机器人中要包含一组试探方式，以避免环路的出现。

总的说来，爬虫的自动化程度越高（人为的监管越少），越可能陷入麻烦之中。机器人的实现者需要做一些取舍——这些试探方式有助于避免问题的出现，但你可能会终止扫描那些看起来可疑的有效内容，因此这种方式也是“有损失”的。 222

在机器人会遇到的各种危险的 Web 中，有些技术的使用可以使机器人有更好的表现。

注 5：这是 http://www.searchtools.com/robots/robot-checklist.html 上提到的日历站点 http://cgi.umbc.edu/cgi-bin/WebEvent/Webevent.cgi 上的真实例子。这样的动态内容带来的后果就是，很多机器人都拒绝爬行 URL 中包含子字符串“cgi”的页面。