器的访问日志杂乱不堪，甚至会使一些比较脆弱的 Web 服务器崩溃。

- *爱打听的机器人*

  有些机器人可能会得到一些指向私有数据的 URL，这样，通过因特网搜索引擎和其他应用程序就可以很方便地访问这些数据了。如果数据的所有者没有主动宣传这些 Web 页面，那么在最好的情况下，他只是会认为机器人的发布行为惹人讨厌，而在最坏的情况下，则会认为这种行为是对隐私的侵犯。[17]

  通常，发生这种情况是由于机器人所跟踪的、指向“私有”内容的超链已经存在了（也就是说，这些内容并不像其所有者认为的那么隐密，或者其所有者忘记删除先前存在的超链了）。偶尔也会因为机器人非常热衷于寻找某站点上的文档而出现这种情况，很可能就是在没有显式超链的情况下去获取某个目录的内容造成的。

  从 Web 上获取大量数据的机器人的实现者们应该清楚，他们的机器人很可能会在某些地方获得敏感的数据——站点的实现者不希望通过因特网能够访问到这些数据。这些敏感数据可能包含密码文件，甚至是信用卡信息。很显然，一旦被指出，就应该有某种机制可以将这些数据丢弃（并从所有搜索索引或归档文件中将其删除），这是非常重要的。现在已知一些恶意使用搜索引擎和归档的用户会利用大型 Web 爬虫来查找内容——有些搜索引擎，比如 Google，[18] 实际上会对它们爬行过的页面进行归档，这样，即使内容被删除了，在一段时间内还是可以找到并访问它。

- *动态网关访问*

  机器人并不总是知道它们访问的是什么内容。机器人可能会获取一个内容来自网关应用程序的 URL。在这种情况下，获取的数据可能会有特殊的目的，计算的开销可能很高。很多 Web 站点管理员并不喜欢那些去请求网关文档的幼稚机器人。

9.4 拒绝机器人访问

机器人社团能够理解机器人访问 Web 站点时可能引发的问题。1994 年，人们提出了一项简单的自愿约束技术，可以将机器人阻挡在不适合它的地方之外，并为网站

注 17：通常，如果某资源可以通过公共因特网获取的话，它很可能会在某处被引用。由于因特网上链路网的存在，很少有资源是真正私有的。

注 18：参见 http://www.google.com 上的搜索结果。已缓存链接就是 Google 爬虫解析并索引过的页面的副本，大多数搜索结果中都会有已缓存链接。