表9-2 拒绝机器人访问标准的版本

版 本	标题及描述	日 期
0.0	拒绝机器人标准——Martijn Koster 提出的带有 Disallow（不允许）指令的原始 robots.txt 机制	1994 年 6 月
1.0	控制 Web 机器人的方法——Martijn Koster 提供了额外支持 Allow（允许）的 IETF 草案	1996 年 11 月
2.0	拒绝机器人访问的扩展标准——Sean Conner 提出的扩展标准，包括了正则表达式和定时信息，没有得到广泛的支持	1996 年 11 月

现在大多数机器人采用的都是标准 v0.0 或 v1.0。版本 v2.0 要复杂得多，没有得到广泛的应用。可能永远也不会得到广泛应用。这里我们重点介绍 v1.0 标准，因为它的应用很广泛，而且与 v0.0 完全兼容。

9.4.2 Web站点和robots.txt文件

如果一个 Web 站点有 robots.txt 文件，那么在访问这个 Web 站点上的任意 URL 之前，机器人都必须获取它并对其进行处理。[19] 由主机名和端口号定义的整个 Web 站点上仅有一个 robots.txt 资源。如果这个站点是虚拟主机，每个虚拟的 docroot 都可以有一个不同的 robots.txt 文件，像所有其他文件一样。

通常不能在 Web 站点上单独的子目录中安装“本地”robots.txt 文件。网管要负责创建一个聚合型 robots.txt 文件，用以描述 Web 站点上所有内容的拒绝访问规则。

1. 获取robots.txt

机器人会用 HTTP 的 GET 方法来获取 robots.txt 资源，就像获取 Web 服务器上所有其他资源一样。如果有 robots.txt 文件的话，服务器会将其放在一个 text/plain 主体中返回。如果服务器以 404 Not Found HTTP 状态码进行响应，机器人就可以认为这个服务器上没有机器人访问限制，它可以请求任意的文件。

机器人应该在 From 首部和 User-Agent 首部中传输标识信息，以帮助站点管理者对机器人的访问进行跟踪，并在站点管理者要查询，或投诉的机器人事件中提供一些联系信息。下面是一个来自商业 Web 机器人的 HTTP 爬虫请求实例：

```
GET /robots.txt HTTP/1.0
Host: www.joes-hardware.com
User-Agent: Slurp/2.0
Date: Wed Oct 3 20:22:48 EST 2001
```

注 19：尽管我们说的是 robots.txt 文件，但 robots.txt 资源并不一定要严格地位于文件系统中。比如，可以由一个网关应用程序动态地生成这个 robots.txt 资源。