本章我们来仔细了解一下被称为 Web 机器人（Web robot）的自活跃（self-animating）用户代理，以继续我们的 HTTP 架构之旅。

Web 机器人是能够在无需人类干预的情况下自动进行一系列 Web 事务处理的软件程序。很多机器人会从一个 Web 站点逛到另一个 Web 站点，获取内容，跟踪超链，并对它们找到的数据进行处理。根据这些机器人自动探查 Web 站点的方式，人们为它们起了一些各具特色的名字，比如“爬虫”、“蜘蛛”、“蠕虫”以及“机器人”等，就好像它们都有自己的头脑一样。

这里有几个 Web 机器人的示例。

- 股票图形机器人每隔几分钟就会向股票市场的服务器发送 HTTP GET，用得到的数据来构建股市价格趋势图。
- Web 统计机器人会收集与万维网规模及发展有关的“统计”信息。它们会在 Web 上游荡，统计页面的数量，记录每个页面的大小、所用语言以及媒体类型。[1]
- 搜索引擎机器人会搜集它们所找到的所有文档，以创建搜索数据库。
- 比较购物机器人会从在线商店的目录中收集 Web 页面，构建商品及其价格的数据库。

9.1 爬虫及爬行方式

Web 爬虫是一种机器人，它们会递归地对各种信息性 Web 站点进行遍历，获取第一个 Web 页面，然后获取那个页面指向的所有 Web 页面，然后是那些页面指向的所有 Web 页面，依此类推。递归地追踪这些 Web 链接的机器人会沿着 HTML 超链
215 创建的网络“爬行”，所以将其称为*爬虫*（crawler）或*蜘蛛*（spider）。

因特网搜索引擎使用爬虫在 Web 上游荡，并把它们碰到的文档全部拉回来。然后对这些文档进行处理，形成一个可搜索的数据库，以便用户查找包含了特定单词的文档。网上有数万亿的 Web 页面需要查找和取回，这些搜索引擎蜘蛛必然是些最复杂的机器人。我们来进一步仔细地看看这些爬虫是怎样工作的。

9.1.1 从哪儿开始：根集

在把饥饿的爬虫放出去之前，需要给它一个起始点。爬虫开始访问的 URL 初始集合被称作*根集*（root set）。挑选根集时，应该从足够多不同的站点中选择 URL，这样，爬遍所有的链接才能最终到达大部分你感兴趣的 Web 页面。

注 1：http://www.netcraft.com 收集了大量统计度量值，用于统计 Web 站点使用的是哪种类型的服务器。