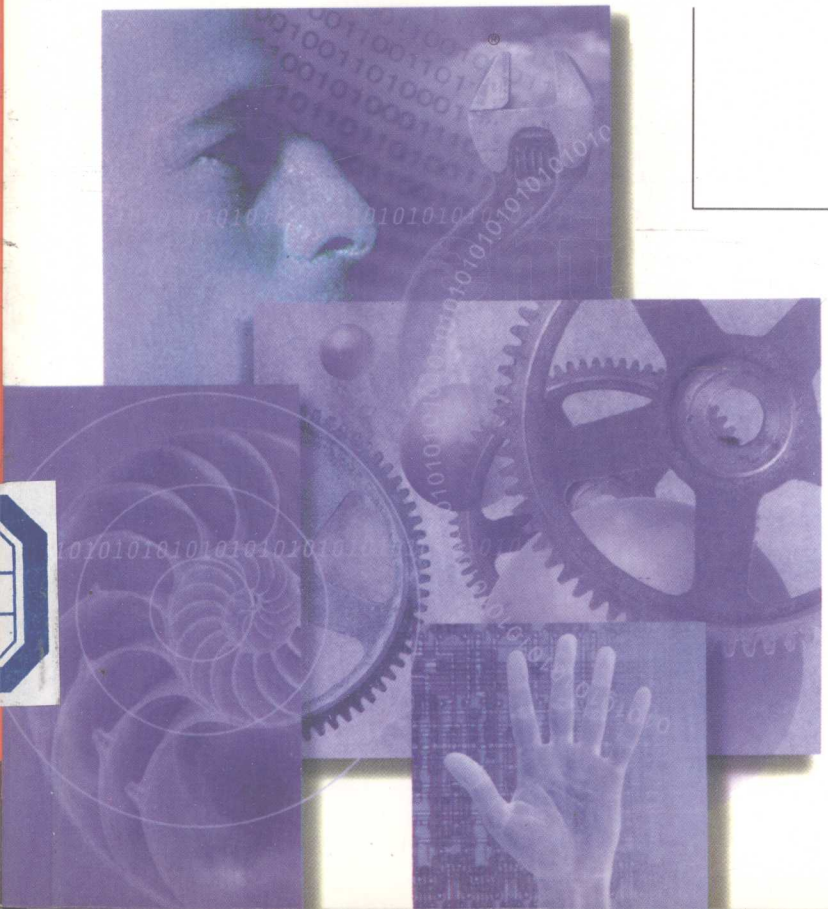
21 BEIJING HOPE

Designed for Microsoft Windows NT Windows 98

Microsoft Visual Studio 中文版系列图书
编程的利器 · 知识的迸发

Programming Bots, Spiders, and Intelligent Agents in Microsoft Visual C++

自动、查询和智能代理程序设计



为 Win32® 平台创建
新一代智能助理

本书配套光盘内容包括:
1. 源代码例子
2. 自动程序类
3. AlphaCONNECT BusinessVue, AlphaCONNECT StockVue
4. Microsoft Internet Explorer 4

[美] David Pallmann 著
希望图书创作室 译

北京希望电子出版社
Beijing Hope Electronic Press
www.bhp.com.cn

Microsoft Press

美国微软出版

开发人员的

Programming Bots, Spiders, and Intelligent Agent in Microsoft Visual C++

Visual C++ 6.0自动、查询和智能代理程序设计

[美] David Pallmann 著

希望图书创作室 译

本书配套光盘内容包括：

1. 源代码例子
2. 自动程序类
3. AlphaCONNECT BusinessVue
   AlphaCONNECT StockVue
4. Microsoft Internet Explorer 4

北京希望电脑公司
北京希望电子出版社
Beijing Hope Electronic Press
www.bhp.com.cn

1999

内 容 提 要

随着 Internet 的迅猛发展，与之相关的软件范畴也逐步得到了成熟和完善。本书详细地讨论了这些内容的基础知识，并着重讲述了如何用 Visual C++和 Microsoft 基础类库（MFC）来实现这些技术。

本书共 20 章，分为五个部分。第一篇讲述了 Bot 的概念，探讨了各种 Bot 程序，描述了访问 Internet 的方法和规则，还讲解了规划自动进程的方法，登录的种类以及用于 Robot 程序的 C++类。第二篇着重讲解一类特殊的 Bot，称之为 Spider，其中介绍了实现探索系统、站点爬行和多线程的技术。第三篇讲述智能代理及能使之有效工作的大量编程组件，其中详细介绍了用户界面设计、解释数据的不同方法及事件、警示和通知。第四篇深入介绍 Bot、Spider 和智能代理中用到的技术。第五篇介绍了书中代码的风格及使用方法。

本书本书的材料组织严密，内容由浅入深，由易到难。本书特别适合用 Visual C++进行 Internet 开发的编程人员参考，也可供 Internet 上开发的初学者、大专院校师生自学、教学参考用书和社会相关领域培训班教材。

本书配套光盘内容包括：1. 源代码例子；2. 自动程序类；3. AlphaCONNECT BusinessVue，AlphaCONNECT StockVue；4. Microsoft Internet Explorer 4。

版 权 声 明

书　　名：Visual C++ 6.0 自动、查询和智能代理程序设计
文本著作者：[美] David Pallmann　著　　希望图书创作室　译
审校/责任编辑：周　艳
CD 制 作 者：希望多媒体开发中心
CD 测 试 者：希望多媒体测试部
出版、发行者：北京希望电脑公司　北京希望电子出版社
地　　址：北京海淀路 82 号，100080
网址：www.bhp.com.cn
E-mail：lwm@hope.com.cn
电话：010-62562329,62541992,62637101,62637102(图书发行,技术支持)
010-62633308,62633309(多媒体发行，技术支持)
010-62613322-215（门市）　010-62531267(编辑部)
经　　销：各地新华书店、软件连锁店
排　　版：希望图书输出中心
CD 生 产 者：文录激光科技有限公司
文本印刷者：北京双青印刷厂
开本/规格：787×1092　16 开本　　42.75 印张　　771 千字
版次/印次：1999 年 10 月第 1 版　　1999 年 10 月第 1 次印刷
印　　数：0001-5000 册
本 版 号：新出音管[1998]312 号　ISBN7-980026-59-4/TP・46
定　　价：66.00 元（1CD 含配套书）

译者序

Internet 是当今世界上最大的信息网络，它遍布全球，庞大无比，正在从根本上改变着世界的面貌和人类的生存方式。Internet 的发展带动了新的软件范畴的发展，包括浏览器、搜索引擎、Bot、Spider 和智能代理等。

本书从编程的角度介绍了 Bot、Spider 和智能代理三大部分内容。详细地讲述了它们是什么、为什么要用到它们、它们如何工作以及如何利用 Visual C++和 Microsoft 基础类库（MFC）编写 Internet 应用程序。对于 Bot，本书不仅探讨了不同种类的 Bot，同时还讨论了关于自动访问 Internet 的一些问题和规则；对于 Spider，本书充分讨论了它的特性和功能，提供了很多相关的乘法，同时还介绍了多线程的实现技术。智能代理使得 Robot 编程技术突飞猛进。在未来的几年内，智能代理将无孔不入地渗入我们的生活。本书以相当大的篇幅讲解智能代理。智能代理要比 Bot 和 Spider 复杂，且已经形成一个专门的技术领域，独自处理复杂的事务，可以自作主块，因而编写智能代理程序既富于挑战性又充满乐趣。

本书以程序类或代码片断相结合的方式，介绍实现 Bot，Spider 和智能代理的方法。大部分程序示例都是与网络相关的。为了照顾那些喜欢迎接挑战的读者，很多程序示例都提供了扩展的方法，读者可自行试验。除了上述程序示例外，书中还扩展了 C++类，以便重复使用，这些内容都包含在配套光盘中。

本书由希望图书创作室策划。由姜旭光、冷冰、季伟敢、张蓓、王江胜、陈莹、孙立军、郭宇、范钟勤、山岗、陈杰、陈世晨、韩永海、郑强、刘海波、邓青云翻译，张小同参加了稿件的整理工作。在本书的翻译和出版过程中，北京希望电子出版社给予了鼎力支持，在此一并谨致衷心感谢。

由于翻译人员水平有限，译文难免会有错误，欢迎读者批评指正。

作者简介

DAVID PALLMANN

David Pallmann 是微软认证的方案开发员，也是 AlphaCONNECT 的发明人。这是智能客户机领域的一项前沿技术。他是位于加州 Santa Ana 的 AlphaServ 公司（前身是 Alpha Microsystems 公司）的技术指导，也是许多商业软件和智能客户机产品，包括 AlphaCONNECT BusinessVue,StockVue 和 Spotlight 的设计者。

David 的家在南加州，与他的妻子 Rebekah 和他们两岁的女儿 Susan 生活在一起。工作之余，David 喜欢和他的家人在一起阅读科技小说，在周末学校义务授课和唱歌。他的近期主要目标是作为一名竞争者克服困难，解决 Y2K。他在互连网上的地址是 www.alphaconnect.com/agents，我们在这里可以找到他。

感谢

由于很多人直接或间接的帮助，本书才得以诞生，向这些人一一表示感谢是不可能的，我只能提到给我最大帮助的那些人。

尽管很早就想到写这本书了。但还是微软公司的 Parri Munsell 促成此事，他一再鼓励我写这本书，把我介绍给微软出版社的负责人，若不是他的鼓励，本书到目前为止可能还只是个构想。

在最后几年里，我有幸与 Alphaserv.com(前身为 Alpha Microsystems)的一批杰出的软件工程师共事，我的知识来自工作组中现在的和以前的成员，以及继续接受教育的信念.Liane Angeles，Christopher Hunter，Katie Huynh，Terry Partridge，John Ross，Jed Stumpf 和 Peter Tran 对本书的完成起了关键作用。在交付微软出版社之前，Chris 和 John 审阅了手稿和代码示例。

很荣幸在 Alphaserv.com 结识了一些前辈：总裁兼首席执行官 Douglas Tullio，分管营销的副总裁 Denny Michael，分管工程及生产的副总裁 John Glade。他们使我自由地追寻新思想，发挥创造力，这使我的工作充满乐趣；他们也全心全意地支持本书的观点。

在编写本书的日子里，妻子 Rebekah 对我非常理解和支持，我们一岁半的女儿 Susan 在爸爸工作时也总努力安静下来。编写本书是一项长期的工作，没有这些支持是不可能完成的。

与微软出版社的合作是非常愉快的，参与编辑技术监督以及印刷工作的所有人都是行家并且充满热情，与我合作最密切的有三个人：首先要感谢编辑 Ben Ryan，他赞成本书的观点，并提出在微软出版社出版；显然，本书离不开他的鼓励。在本书出版期间，Michelle Goodman 既是项目编辑又是手稿编辑，她在这两方面都很优秀，她既是个行家，但又不失礼节甚至偶尔批驳，如果你觉得本书有一定的可读性，那是她的功劳。作为技术编辑，Donnie Cameron 起到了关键作用：从专业术语到程序代码，他都仔细作了校订。他还使得本书在编辑体例专业化的同时又兼顾了原有的技术内涵和语言的准确性。以后再出版书时，我还会毫不犹豫地与这个工作组合作。

最后还要感谢我的父母，他们为我付出了很多。谢谢诸位。

前言

公众对 Internet 的厚爱引起了软件方面的很多变化和发展，新的软件范畴发展起来，包括浏览器、搜索引擎、bot、spider 和智能代理。浏览器和搜索引擎要与人发生交互作用，而 bot、spider 和智能代理则具有自主性。

本书介绍了浏览器、搜索引擎、bot、spider 以及智能代理的基础知识：它们是什么，人们为什么要用到它们，它们如何工作，如何用 Visual C++和 Microsoft 基础类（MFC）编写这些程序。还讲了如何设计、执行高水平的有商业价值、高效、专业化的 bot 和智能代理。开发平台为 32 位 Microsoft Windows——即 Windows 95（或更高版本)和 Windows NT 4.0（或更高版本）。

Robot 程序是 Internet 上的幕后英雄，但是这种情况将会改变，以前，bot 只是用于执行一些特定的任务，如为搜索引擎建立数据库，如今被称作智能代理的新一代智能型助手的诞生，使得 Robot 编程技术突飞猛进。几年之内，智能代理将无孔不入地渗入我们的生活。

本书以相当大的篇幅讲解智能代理，因为它们生性复杂，不像简单的 bot 和 Spider，从很多角度考虑，智能代理都是软件中最有趣的一类：它们形成一个专门的技术领域，独自进行复杂的工作，可以自作主张，了解它们所代表的人的想法，代理人有他们自己的个性，因而编这种程序既富于挑战性又充满乐趣。

本书的每一章都以程序、类或代码片断结束，每个程序例子都是一个丰富的项目，用它可以建立实现一定功能的 bot，Spider 或智能代理。大部分程序示例都是与网络相关的 bot，为照顾那些喜欢迎接挑战的读者，很多程序示例都提供了扩展的方法，读者可自行试验。除了这些程序示例，很多章节扩展了 C++类，以便于重复使用，配套光盘中包含所有这些类和程序示例。

尽管本书是面向那些 MFC 程序员的，但是要用别的语言，例如 Java 或 Visual Basic，来编写 bot 或智能代理，本书同样有用。书中总结出的一些原理则同样适用于别的开发环境和编程语言，而且，如果只想知道 bot 和智能代理是如何工作的，可以跳过代码部分，本书依然是有价值的。

本书共二十章，各章有很强的独立性，因此，即使跳过几章，也能理解所讨论的内容，每章都能让你很快编写代码，开发可反复使用的类，让你学的很快。

本书分为以下五部分：

- 第一篇，本篇讲述了 bot 的概念，探讨了各种 bot 程序，描述了访问 Internet 的方法和规则，还讲解了规划自动进程的方法，登陆的种类以及用于 Robot 程序的 C++类。

- 第二篇，这一部分着重于讲解一类特殊的 bot，称之为 Spider。其中介绍了实现探索系统、站点爬行和多线程的技术。
- 第三篇，这一部分讨论智能代理及使之有效工作的大量编程组件，其中详细介绍了用户界面设计、解释数据的不同方法及事件、警示和通知。
- 第四篇，本篇深入介绍 bot、spider 和智能代理中用到的技术，包括访问 Internet、访问数据、传递电子信息等。Bot 不能凭空进行操作，它必须与系统、数据或人以多种方式交互作用。
- 第五篇，书中有很多代码，因而代码风格和变量的命名规则不容忽视，在这个问题上不可能获得一致意见，因此我选了一种比较折中的办法，目的是要让代码易读、易学。

书中所用代码风格是在我个人风格的基础上修订而成的，以便付诸印刷时有更好的可读性。左括号和右括号各单独占用一行，注释如//End 跟在右括号之后，这样看起来结构比较清晰，我还使用了一些常用的缩排，例如：

```
// Repeatedly process all tasks until canceled
while (!bCanceled)
{
    for (int nCurrTask = 0; nCurrTask < nTasks; nCurrTask++)
    {
        AllocateTask();
        ProcessTask(nCurrTask);
        DeallocateTask();
    } // End for
    Sleep(100);
} // End while
```

变量名又是一个问题，我喜欢匈牙利标记法，但不喜欢使用很长的变量名，那些无论是输入还是读起来都很麻烦，既然代码都不大，我就使用匈牙利标记法，以 n 打头表示整型量，如 nCount，以 s 打头表示 CString 型，如 sName，b 打头表示布尔变量，如 bInitialized，p 打头表示指针，如 pTray。

当定义一个重要的类成员时，我让变量名以 m-打头，如 m-nCount，但是对于那些只有几个变量的小对话框则没必要这么复杂。定义一个全局变量时，以 g_打头，如 g_cBuffer，我一般只在较大的程序中使用 g_前缀。

当声明一个与控件有关的对话框类的成员变量时，我使用 m_前缀，变量名与资源 ID 相同，如一个控件的资源 ID 为 IDC_NAME，它的变量名称就是 m_name，某成员变量表示一个控件而非一个值，我就在名称结尾加上_ctl，如 m_name_ctl。

目　录

第一篇　Bot

第二篇　Spider

第三篇　智能代理

第四篇 潜在的技术

第五篇 附录

第一篇 Bot

第一章介绍了 Robot 程序的概念,探讨了不同种类的 bot，并讨论了关于自动访问 Internet 的一些问题和规则,这一章的程序示例判断一个 bot 是否有资格访问 web 服务器。

第二章讲述了 Internet 通信中的一些关键概念,以及通用的协议和 HTML 及 XML 语言,这一章对浏览器进行了细致入微的分析。

第三章介绍了本书中程序示例所用到的 C++类，当你在编程时用到这些类时，可以这一章作为参考资料。

第四章介绍了使 Robot 定时进行后台操作的方法，这段被称作 webwatch 的代码周期性地检测一系列网址，一有变化就立即报告。

第五章介绍了 Robot 是如何记录它们成功或失败的足迹的，这段 webmonitor 程序反复检测一系列 web 服务器，看他们是否有响应，并将它的发现记入日志。

第一篇 Bot

第一章　Bots:Internet 的幕后英雄

Asimov 的 Robot 三原则：

1. Robot 不会伤害人类，或是由于其无所事事而使人类受损失。
2. 在不违反原则 1 的前提下，Robot 必须遵守人类给它发布的任何命令。
3. 只要不违反前两条原则，Robot 应保护自己永远生存下去。

这一章将使你认识 bots，或者说是 Robot 或 Robot 程序，你将学到 bot 的大量应用，以及应用程序访问 Internet 时会出现的问题，另外，本章还告诉你如何确保你的 bot 成为良好的线上公民，一旦认识到 bots 是如何行使其职责的，你就能用 Microsoft Visual C++编写你的第一个 bot 了。

究竟什么是 bot

首先要了解 bot spider 与智能代理的区别，这一点非常重要，网络编程往往是不同的人面对不同的对象，尤其是当他们要用到还处于发展中的技术时。

所有用于后台操作的程序都可以叫作 bot，尽管这个词在访问 Internet 时常用于后台程序。根据这个定义，Spider 和智能代理都是 bot 家族中的成员。事实上，你可以在 web 上看到各种不同的 bot，诸如：Infobot，Newsbot 以及 WebBot。

Spider（网上爬行者）是对专门的网址，有时甚至是整个环球网进行索引，映射的 bot，他们使得搜索引擎成为可能。听说过 HotBot 搜索引擎吗？现在你知道了它名字的由来，本书的第二部分将对 Spider 进行更详细的讨论。

智能代理.在技术上也属于 Bot，但它们又足以自成一派，这有两个原因：首先，术语“bot”(尤其是老的叫法“robot”）总让人将其与“愚蠢”“笨拙”等联系到一块，而智能代理却绝非这样简单；其次，尽管智能代理常常在不知不觉中工作着，它们也有交互式部件，可以接收指令，或向用户汇报任务的结果，本书第三部分将对智能代理进行深入探讨。

现实世界中的 Bot

在进一步讨论之前，我们先看几个 bot。下面要讲的第一个例子是一个搜索引擎，

它借助于 Spider 来建立当前网址数据库；第二个例子描述了一个智能代理对新闻进行临控，当某一主题有新的信息出现，立即向用户汇报。

Spider 示例：搜索引擎

当你使用 web 浏览器来访问一个搜索引擎时，将看到的只是引擎中被称作“前端”的一半景象。它是这样的：浏览器与 web 服务器相连，服务器弹出一个窗口让你输入要搜索的条文。当你提出搜索请求，web 服务器就在一个大型数据库中，查找相匹配的条文，列出找到的站点，把结果传给浏览器，图 1.1 是搜索引擎的前端示意图。

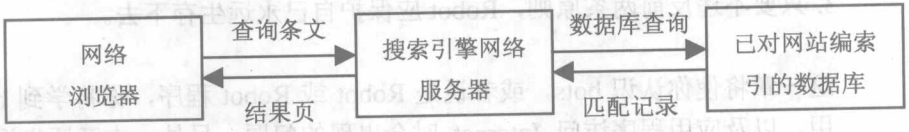


图 1.1　搜索引擎前端

你所看不见的另一半景象就是“Spider”，或“网络爬行者”，这段程序速度扫描 Internet 并更新数据库，系统的这部分根本无法看到，因而被称作搜索引擎的“后端”。Spider 和引擎的前端一样重要，如果没有它，搜索引擎的网址数据库很快就会过，图 1.2 是引擎的后端示意图。

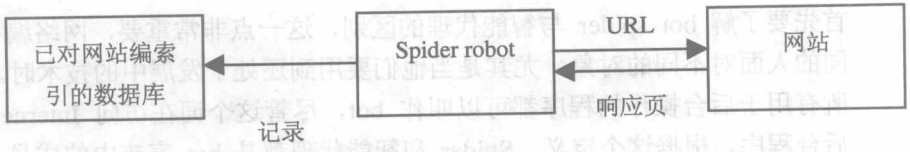


图 1.2　搜索引擎后端

Spider 必须不停地扫描 Internet，对遇到的网址进行索引，这意味着 Spider 必须浏览主页的每个结点——这一过程叫作爬行。

大多数搜索引擎都让 Spider 来做这项工作，例如一些流行的实时搜索引擎 Altavista，Excite webCrawler，Lycos 以及 Infoseek。

智能代理示例：新闻 Bot

智能代理在后台服务，新闻 bot 就是一种智能代理，它不停地监控线上的新闻，并为你提供一些有趣的信息，这种助手有着非常重要的价值。网上新闻太多，你不可能全看一遍，但是很可能有些你喜欢的主题。新闻 bot 在后面默默地工作着，发现什么有价值的新闻就会唤起你的注意。

与别的 bot 不同，象新闻 bot 这样的智能代理与用户保持联系，用户需要提出他感兴趣的主题并可能随时进行修改，新闻 bot 必须把符合要求的新闻告诉用户，

图 1.3 表现了一个新闻 bot 的典型结构。

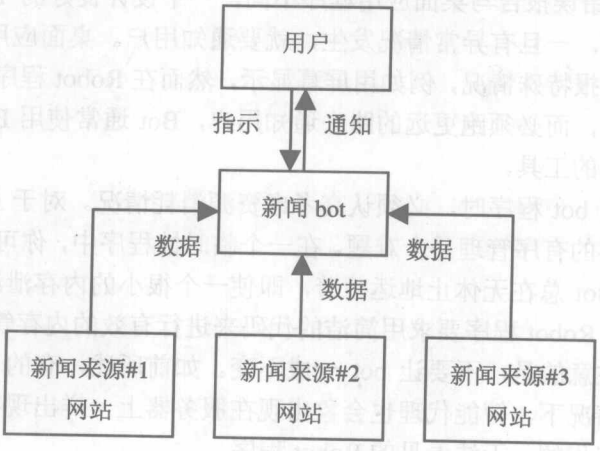


图 1.3 典型的新闻 bot 结构

现实世界中的新闻 bot 包括 BusinessVue 和 StockVue，二者都是 AlphaCONNECT 公司的免费产品，BusinessVue 追踪与公司有关的信息，包括新闻、档案和研讨会，StockVue 追踪关于股票和合作基金的消息，包括新闻和股票报价，在本书的配套光盘中可以看到这两种产品。

Spider 和智能代理的差异

以上两个 bot 示例大不相同。搜索引擎中的 Spider 几乎是看不见的，事实上，大多数人甚至都感觉不到它的存在，尽管新闻 bot 也在我们的视野之外，但它偶尔会与用户发生交互；搜索引擎的 Spider 维护并更新一个服务于大量用户的数据库，而新闻 bot（智能代理）为私人所有，为单个的用户服务；Spider 只在搜索引擎设备的服务器上运行，而智能代理是在用户系统上运行。

尽管 Spider 和智能代理可能访问同一站点，但有个最大的区别：Spider 只对网址进行索引，而智能代理则使用网站中的数据。

尽管二者有所差异，但它们在内部的工作极为相似：它们都需访问 Internet，都在默默无闻地工作着，都对网络服务器的响应进行分析并根据它们所得到的信息做出判断，再把结果发送到别的地方。

后台程序

现在应该明白，bot 生来就具有很低的可见性：它们总是以自己的方式工作着，无需用户交互，往往是在服务器上工作而不是在桌面系统上。

由于 bot 长期在后台操作，它们需要特殊的关照，尤其要注意它们的错误报告和

资源消耗情况。

bot 的错误报告与桌面应用程序不同，一个设计良好的 bot 在一切正常时应当默不作声，一旦有异常情况发生，就要通知用户。桌面应用程序通过视觉方式来向用户通报特殊情况，例如用屏幕显示，然而在 Robot 程序中，别指望用户会在 PC 机面前，而必须跑更远的路去通知用户，Bot 通常使用 E-mail 和工作日志作为错误通报的工具。

在编写 bot 程序时，必须认真考虑资源消耗情况。对于 C++的编程者来说，要做到内存的有序管理是个难题。在一个临时性程序中，你可以忽略清除内存的问题，但是 Bot 总在无休止地运行着，即使一个很小的内存泄漏最终也会导致系统资源耗尽，Robot 程序要求用简洁的代码来进行有效的内存管理。

需要注意的是，不要让 bot 一成不变。如前所述，有的 bot 确实出现在桌面上；特殊情况下，智能代理也会象出现在服务器上一样出现在桌面上。本部分探讨在后台工作的，无法看见的 Robot 程序。

Bot 与 Internet

大多数 bot 以特定的方式浏览 Internet，它们的行为受到网站管理器的监视。和人一样，不同的 bot 有不同的声誉，有的 bot 因其良好的作风受到普遍的尊敬，而有的则是肇事者，谁都躲着它。如果想让你的 bot 生命力持久，就得让它们尊敬 Internet 的内部组织，遵守其法规。

锤打

大多数网站都面向人类，而非 bot，如果 bot 向某一站点妥协的话，就要承担激怒网站管理员的危险。当一个站点因 bot 性能低劣而难以得到访问时，就应当作出妥协。频繁地寻找某一站点被叫作“锤打”，这是使你的程序引起注意的主要方法——也是错误的方法。

最常见的三种锤打是：隔离式锤打，聚合式锤打和分布式锤打，一个程度反复访问同一站点就称作隔离式锤打，这会增加大量的通道，直至回溯到系统的 Internet 协议（IP）地址，而 bot 就是从那里开始运行的。

聚合式锤打源于一个多线程的 bot，或者说，是由于同一 Robot 程序的几个实例同时运行引起的。当对某一站点的访问过于频繁，出乎 bot 编程者的预料，聚合式锤打便发生了。有时为了追踪越来越多的站点，不得不同时运行越来越多的 bot，这时，锤打就会不知不觉地发生。

如果你的 Robot 程序处在公共场所，很多人都在经常使用，那么，当用户多到一