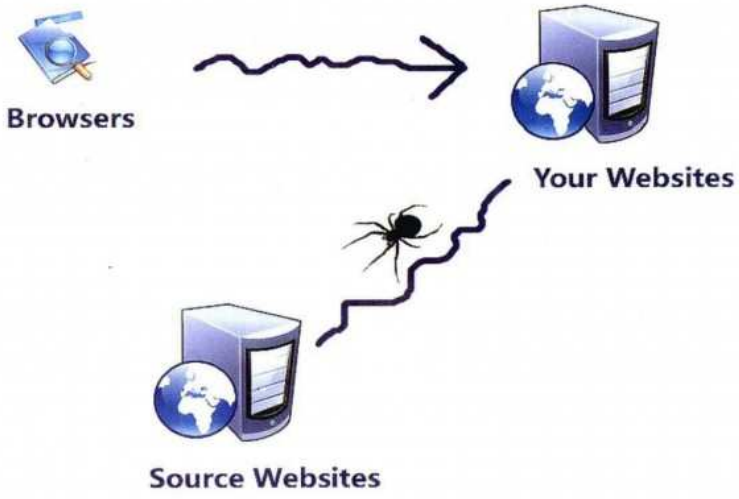
国内**第一本**专门讲解**网络爬虫**开发的书籍
介绍如何应用**云计算**架构开发分布式爬虫



罗 刚 王振东 编著

自己动手写网络爬虫

- 猎兔搜索工程师多年项目经验总结
- 深入介绍Web数据挖掘实现过程
- 光盘中提供了高效的代码解决方案
- 案例均使用流行的Java语言编写

清华大学出版社

自己动手写

网络爬虫

本书介绍了网络爬虫开发中的关键问题与Java实现，主要包括从互联网获取信息、提取信息和对Web信息挖掘等内容，本书在介绍基本原理的同时注重通过具体代码实现来深入理解。书中部分代码甚至可以直接应用。本书适用于有Java程序设计基础的开发人员，同时也可以作为计算机相关专业本科生或研究生的参考教材。

ISBN 978-7-302-23647-4

9 787302 236474 >

定价：43.00元（附光盘1张）

前　言

当你在网上冲浪时，你是否知道还有一类特殊的网络用户也在互联网上默默地工作着，它们就是网络爬虫。这些网络爬虫按照设计者预定的方式，在网络中穿梭，同时自动收集有用的信息，进行分类和整理，将整理结果提供给用户，以方便用户查找他们感兴趣的内容。由于网络爬虫的实用性，引起了很多程序员，特别是 Web 程序员的兴趣。

但是大多数网络爬虫的开发原理与技巧在专业的公司内部都秘而不宣，至今仍然缺少理论与实践相结合的专门介绍网络爬虫的书籍。本书将弥补这个问题，尝试理论与实践相结合，深入透彻地讲解网络爬虫的原理，并且辅以相关代码作为参考。本书相关的代码在附带光盘中可以找到。

本书的两位主要作者在搜索引擎领域都有丰富的理论和实践经验。同时，还有多个程序员帮忙开发或编写了代码实现，例如 Java 实现异步 I/O 或对 PDF 文件的处理等。由于作者的日常工作繁忙，做得不够的地方敬请谅解。

作者罗刚在参加编写本书之前，还独立撰写过《自己动手写搜索引擎》一书，但存在讲解不够细致、知识点不够深入等问题。此次与王振东合著本书，相对于上一本书而言，对读者反馈有更高的预期。因为作者相信如下的假设：如果能够与更多的人更好地合作，事情往往能做得更好。

本书从基本的爬虫原理开始讲解，通过介绍优先级队列、宽度优先搜索等内容引领读者入门；之后根据当前风起云涌的云计算热潮，重点讲述了云计算的相关内容及其在爬虫中的应用，以及带偏好的爬虫、信息抽取、链接分析等内容；为了能够让读者更深入地了解爬虫，本书在最后两章还介绍了有关爬虫的数据挖掘的内容。

由于搜索引擎相关领域也正在快速发展中，而且由于篇幅的限制，有些不成熟的内容，没有能够在本书体现，例如有关“暗网”的内容。随着技术的不断发展，我们将在今后的版本中加入这些内容。

本书适合需要具体实现搜索引擎的程序员使用，对于信息检索等相关研究人员也有一定的参考价值，同时猎兔搜索技术团队也已经开发出以本书为基础的专门培训课程和商业软件。目前搜索引擎开发人员仍然很稀缺，作者真诚地希望通过本书把读者带入搜索引擎开发的大门并认识更多的朋友。

感谢开源软件和我们的家人、关心我们的老师和朋友、创业伙伴以及选择猎兔搜索软件的客户多年来的支持。

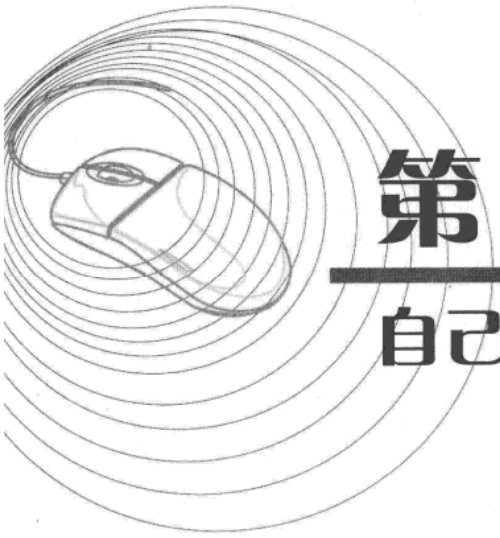
编　者

目　录

第 1 篇　自己动手抓取数据

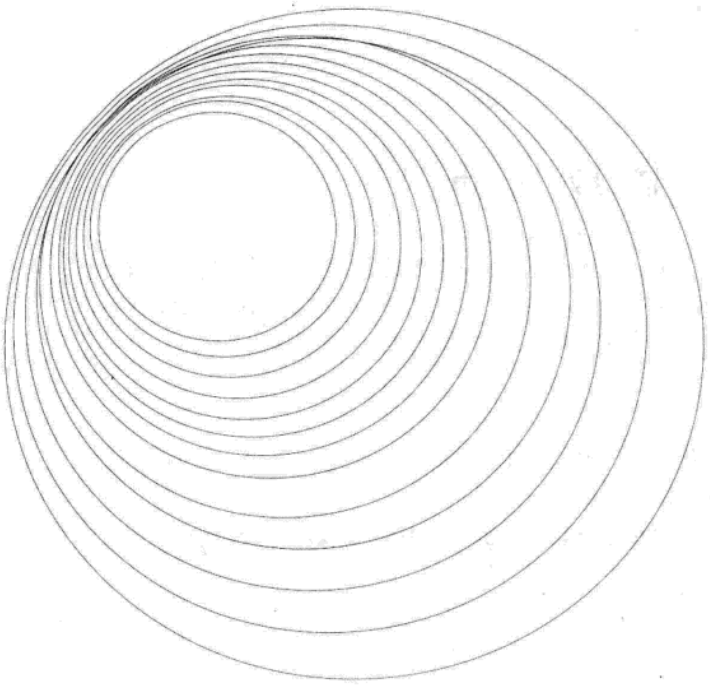
第 2 篇　自己动手抽取 Web 内容

第 3 篇　自己动手挖掘 Web 数据



第 1 篇

自己动手抓取数据

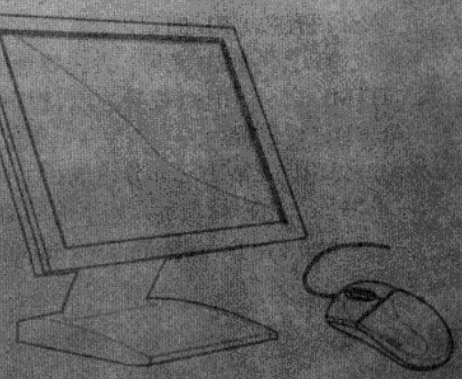


第1章

全面剖析网络爬虫

你知道百度、Google 是如何获取数以亿计的网页并且实时更新的吗？你知道在搜索引擎领域人们常说的 Spider 是什么吗？本章将全面介绍网络爬虫的方方面面。读完之后，你将完全有能力自己写一个网络爬虫，随意抓取互联网上任何感兴趣的东西。

既然百度、Google 这些搜索引擎巨头已经帮我们抓取了互联网上的大部分信息，为什么还要自己写爬虫呢？因为深入整合信息的需求是广泛存在的。在企业中，爬虫抓取下来的信息可以作为数据仓库多维展现的数据源，也可以作为数据挖掘的来源。甚至有人为了炒股，专门抓取股票信息。既然从美国中情局到普通老百姓都需要，那还等什么，让我们快开始吧。



1.1 抓 取 网 页

网络爬虫的基本操作是抓取网页。那么如何才能随心所欲地获得自己想要的页面？这一节将从 URL 开始讲起，然后告诉大家如何抓取网页，并给出一个使用 Java 语言抓取网页的例子。最后，要讲一讲抓取过程中的一个重要问题：如何处理 HTTP 状态码。

1.1.1 深入理解 URL

抓取网页的过程其实和读者平时使用 IE 浏览器浏览网页的道理是一样的。比如，你打开一个浏览器，输入猎兔搜索网站的地址，如图 1.1 所示。

图 1.1 使用浏览器浏览网页

“打开”网页的过程其实就是浏览器作为一个浏览的“客户端”，向服务器端发送了一次请求，把服务器端的文件“抓”到本地，再进行解释、展现。更进一步，可以通过浏览器端查看“抓取”过来的文件源代码。选择“查看”|“源文件”命令，就会出现从服务器上“抓取”下来的文件的源代码，如图 1.2 所示。

在上面的例子中，我们在浏览器的地址栏中输入的字符串叫做 URL。那么，什么是 URL 呢？直观地讲，URL 就是在浏览器端输入的 http://www.lietu.com 这个字符串。下面我们深入介绍有关 URL 的知识。

在理解 URL 之前，首先要理解 URI 的概念。什么是 URI？Web 上每种可用的资源，如 HTML 文档、图像、视频片段、程序等都由一个通用资源标志符(Universal Resource Identifier，URI)进行定位。

URI 通常由三部分组成：①访问资源的命名机制；②存放资源的主机名；③资源自身的名称，由路径表示。如下面的 URI：

http://www.webmonkey.com.cn/html/html40/

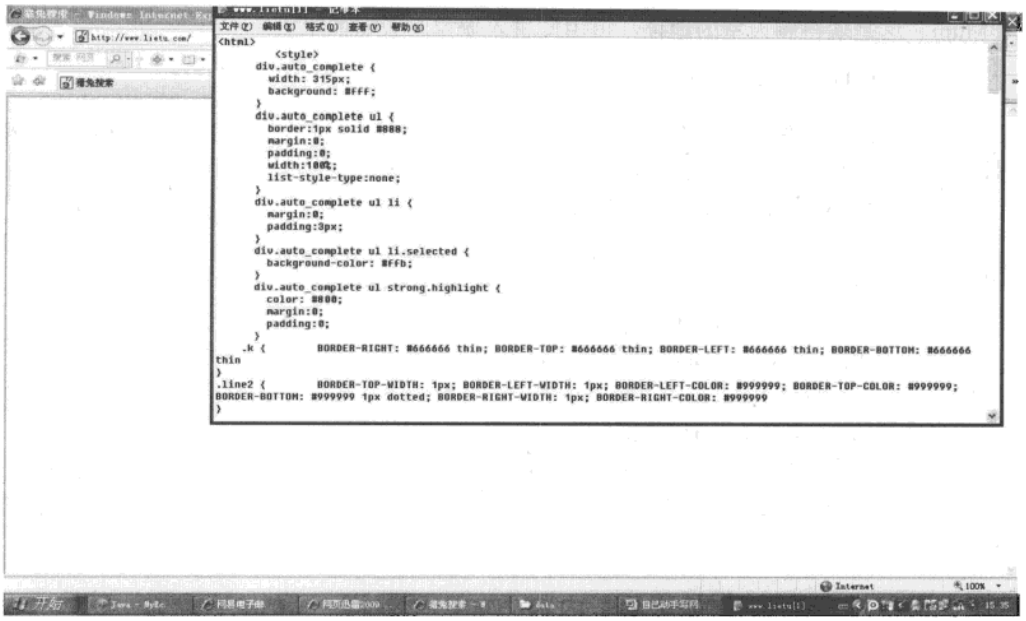


图 1.2　浏览器端源代码

我们可以这样解释它：这是一个可以通过 HTTP 协议访问的资源，位于主机 www.webmonkey.com.cn 上，通过路径“/html/html40”访问。

URL 是 URI 的一个子集。它是 Uniform Resource Locator 的缩写，译为“统一资源定位符”。通俗地说，URL 是 Internet 上描述信息资源的字符串，主要用在各种 WWW 客户程序和服务器程序上，特别是著名的 Mosaic。采用 URL 可以用一种统一的格式来描述各种信息资源，包括文件、服务器的地址和目录等。URL 的格式由三部分组成：

- 第一部分是协议(或称为服务方式)。
- 第二部分是存有该资源的主机 IP 地址(有时也包括端口号)。
- 第三部分是主机资源的具体地址，如目录和文件名等。

第一部分和第二部分用“://”符号隔开，第二部分和第三部分用“/”符号隔开。第一部分和第二部分是不可缺少的，第三部分有时可以省略。

根据 URL 的定义，我们给出了常用的两种 URL 协议的例子，供大家参考。

1. HTTP 协议的 URL 示例

使用超级文本传输协议 HTTP，提供超级文本信息服务的资源。

例：http://www.peopledaily.com.cn/channel/welcome.htm

其计算机域名为 www.peopledaily.com.cn。超级文本文件(文件类型为.html)是在目录/channel 下的 welcome.htm。这是中国人民日报的一台计算机。

例：http://www.rol.cn.net/talk/talk1.htm

其计算机域名为 www.rol.cn.net。超级文本文件(文件类型为.html)是在目录/talk 下的