Broadview®
www.broadview.com.cn

# 网络爬虫全解析

## 技术、原理与实践

罗刚◎著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

# 网络爬虫全解析

## 技术、原理与实践

罗刚◎著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

## 内 容 简 介

本书介绍了如何开发网络爬虫。内容主要包括开发网络爬虫所需要的 Java 语法基础和网络爬虫的工作原理，如何使用开源组件 HttpClient 和爬虫框架 Crawler4j 抓取网页信息，以及针对抓取到的文本进行有效信息的提取。为了扩展抓取能力，本书介绍了实现分布式网络爬虫的关键技术。

另外，本书介绍了从图像和语音等多媒体格式文件中提取文本信息，以及如何使用大数据技术存储抓取到的信息。最后，以实战为例，介绍了如何抓取微信和微博，以及在电商、医药、金融等领域的案例应用。其中，电商领域的应用介绍了使用网络爬虫抓取商品信息入库到网上商店的数据库表。医药领域的案例介绍了抓取 PubMed 医药论文库。金融领域的案例介绍了抓取股票信息，以及从年报 PDF 文档中提取表格等。

本书适用于对开发信息采集软件感兴趣的自学者。也可以供有 Java 或程序设计基础的开发人员参考。



**图书在版编目（CIP）数据**

网络爬虫全解析：技术、原理与实践 / 罗刚著. —北京：电子工业出版社，2017.3
ISBN 978-7-121-31071-3

Ⅰ. ①网… Ⅱ. ①罗… Ⅲ. ①软件工具－程序设计 Ⅳ. ①TP311.561

中国版本图书馆 CIP 数据核字（2017）第 047570 号

策划编辑：董 英
责任编辑：徐津平
印 刷：北京京师印务有限公司
装 订：北京京师印务有限公司
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编：100036
开 本：787×980 1/16 印张：27.75 字数：585 千字
版 次：2017 年 3 月第 1 版
印 次：2017 年 3 月第 1 次印刷
印 数：3000 册 定价：79.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）51260888-819，faq@phei.com.cn。

# 前　言

现代社会，有效信息对人来说就像氧气一样不可或缺。互联网让有效信息的收集工作变得更容易。当你在网上冲浪时，网络爬虫也在网络中穿梭，自动收集互联网上有用的信息。

自动收集和筛选信息的网络爬虫让有效信息的流动性增强，让我们更加高效地获取信息。随着越来越多的信息显现于网络，网络爬虫也越来越有用。

各行业都离不开对信息的采集和加工处理。例如，农业需要抓取气象数据、农产品行情数据等实现精准农业。机械行业需要抓取零件、图纸信息作为设计参考。医药行业需要抓取一些疾病的治疗方法信息。金融行业需要抓取上市公司基本面和技术面等相关信息作为股市涨跌的参考，例如，太钢生产出圆珠笔头，导致它的股票“太钢不锈”上涨。此外，金融行业也需要抓取股民对市场的参与度，作为市场大势判断的依据。

每个人都可以用网络爬虫技术获得更好的生存策略，避免一些糟糕的情况出现，让自己生活得更加幸福和快乐。例如，网络爬虫可以收集到二甲双胍等可能抗衰老的药物，从而让人生活得更加健康。

本书的很多内容来源于搜索引擎、自然语言处理、金融等领域的项目开发和教学实践。感谢开源软件的开发者们，他们无私的工作丰富了本书的内容。

本书从开发网络爬虫所需要的 Java 语法开始讲解，然后介绍基本的爬虫原理。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，以及信息抽取、链接分析等内容。接下来介绍了有关爬虫的 Web 数据挖掘等内容。为了让读者更深入地了解爬虫的实际应用，最后一章是案

例分析。本书相关的代码在读者 QQ 群（294737705）的共享文件中可以找到。

本书适合需要具体实现网络爬虫的程序员使用，对于信息检索等相关领域的研究人员也有一定的参考价值，同时猎兔搜索技术团队已经开发出以本书为基础的专门培训课程和商业软件。目前的一些网络爬虫软件仍有很多功能有待完善，作者真诚地希望通过本书把读者带入网络爬虫开发的大门并认识更多的朋友。

感谢早期合著者、合作伙伴、员工、学员、家人的支持，他们给我们提供了良好的工作基础，这是一个持久可用的工作基础。在将来，希望我们的网络爬虫代码和技术能够像植物一样快速生长。

参与本书编写的还有崔智杰、石天盈、张继红、张进威、刘宇、何淑琴、任通通、高丹丹、徐友峰、孙宽，在此一并表示感谢。

罗　刚

2017 年 2 月

---



# 目　录

# 1

# 第 1 章

# 技术基础

很多种编程语言都可以用来开发爬虫。相对于 Python，Java 由于严谨的语法结构和体系结构，所以在开发爬虫方面有后发优势。

很多网络爬虫是使用 Java 或者 C#语言开发的。如果是开发采集器那样的客户端爬虫，那么可以使用 C#开发爬虫。如果是运行在服务器端的爬虫，则可以用 Java 开发。

只要有目标，你可以做到很多从来没有做过的事情。没有基础也可以学习开发网络爬虫，本章是专门为开发爬虫写的 Java 基础介绍。

## 1.1 第一个程序

Java 程序都运行在虚拟机上。为什么要用虚拟机，而不是直接运行在本机的操作系统上？因为 Windows 是收费的，而 Linux 可以免费使用。可以把 Windows 当作开发环境使用，而把程序部署在 Linux 上。因为运行在指令集相同的虚拟机上，所以 Java 程序可以不经修改地在不同

操作系统之间切换。

并不一定要自己买房子以后才有地方住。并不一定要在本机安装开发环境以后，才能运行第一个 Java 程序。有一些在线的开发环境可运行 Java 程序，例如 http://ideone.com/。

第一个 Java 程序是从一个类中定义的 main 方法开始执行的。

```
public class Crawler{
        public static void main (String args[]) {
            System.out.println("Hello Crawler!");
        }
}
```

底层到底做了些什么？源代码定义了一个叫作 Crawler 的类，虚拟机执行其中的 main 方法。

## 1.2 准备开发环境

Eclipse 也是使用 Java 开发的，所以先准备基本的 Java 开发环境（简称 JDK），然后准备运行在 JDK 上的 Eclipse。

### 1.2.1 JDK

JDK 可以从 Java 官方网站 http://java.sun.com 下载得到。注意，不是从 http://www.java.com 下的 Java 虚拟机。

下载 Java SE，也就是标准版本。Latest Release 是最新发布的安装程序。因为可以在 Windows 或 Linux 等多种操作系统环境下开发 Java 程序，所以有多个操作系统的 JDK 版本供选择。

因为 JDK 是有版权的，所以需要接受许可协议（Accept License Agreement）后才能下载。下载完毕后，使用默认方式安装 JDK 即可。JDK 相关的文件都放在一个叫作 JAVA_HOME 的根目录下。JDK 根目录的命名格式是 C:\Program Files\Java\jdk1.6.0_<version>，最后以一个数字类型的版本号结尾，例如 10 或者 21。

因为一台机器可以安装多个 JDK 和 JVM，为了避免混乱，可以新增环境变量 JAVA_HOME，指定一个默认使用的 JDK。

使用 echo 命令检查环境变量 JAVA_HOME。

```
>echo %JAVA_HOME%
C:\Program Files\Java\jdk1.6.0_10
```

如果只需要使用集成开发环境，配置 JAVA_HOME 环境变量就可以了。为了检查 JAVA_HOME 是否已经正确设置，在任何路径输入 Java 命令“>java -version”显示虚拟机的版本号就可以了。

```
java version "1.6.0_10-rc"
Java(TM) SE Runtime Environment (build 1.6.0_10-rc-b28)
Java HotSpot(TM) Client VM (build 11.0-b15, mixed mode, sharing)
```

如果还需要在控制台下执行，则需要访问编译程序的 javac.exe 或者执行 Java 类的 java.exe。环境变量 PATH 指定了从哪里找 java.exe 这样的可执行文件。可以从多个路径查找可执行文件，这些路径以分号隔开。如果想在命令行运行 Java 程序，还可以修改已有的环境变量 PATH，增加 Java 程序所在的路径，例如 C:\Program Files\Java\jdk1.6.0_10\bin。

然后检查环境变量 PATH。

```
>echo %PATH%
```

为了检查 PATH 是否已经正确设置，在任何路径输入 javac 命令显示 javac 的用法就可以了，也可以用第一个 Java 程序试验下。

```
>javac Crawler.java
>java Crawler
```

运行看是否显示“Hello Crawler!”。

## 1.2.2　Eclipse

就好像理发有推子等专门的理发用具，开发软件也有专门的集成开发环境。开发 Java 程序最流行的工具叫作 Eclipse（http://www.eclipse.org）。

Eclipse 也有很多版本，可以选择最简单的一个版本——Eclipse IDE for Java Developers。Eclipse 是绿色软件，无须安装，解压后就可以直接使用。在 Windows 下，双击就可以解压文件。如果需要专门的解压软件，推荐使用 7z（http://www.7-zip.org/）。

Eclipse 默认是英文界面，如果习惯用中文界面，可以从 http://www.eclipse.org/babel/

downloads.php 上下载支持中文的语言包。

Eclipse 把软件按项目管理，每个项目都有自己的.classpath 文件，指定了源代码路径、编译后输出文件的路径以及这个项目引用的 jar 包的路径。

## 1.3 类和对象

世界上有各种各样的生物，每个生物都属于某一个物种。蜘蛛是一个物种，每个蜘蛛都是一个对象。同样，Java 虚拟机的内存中也有很多各种各样的对象。

使用类可以创建具有相同结构和行为的对象。打印 Hello World 的例子并没有创建对象，因为 main 是一个静态方法，不属于任何一个类。现在创建一个属于对象的方法。

```
public class Spider{
      public void hello(){
         System.out.println("Hello World!");
      }
         public static void main (String args[]) {
         Spider p = new Spider(); //新建一个对象
         p.hello(); //调用 hello 方法
         }
}
```

Java 源文件的扩展名为.java，而且必须与类名相同。上面这个 Spider 类必须放在 Spider.java 文件中。

每个人都由不同的原子构成。每个对象都占据不同的内存空间。使用关键字 new 来为对象分配空间，就是实例化对象。关键字 new 声明了对象的诞生，但是不是所有的数据类型都是对象。一些基本的数据类型，例如 int、boolean 等都不是对象，不能用 new 的方式实例化。

toString 方法返回一个描述对象内部状态的字符串。所有的对象都有 toString 方法，这些共同的方法在 Object 类中定义，Object 是所有对象的共同祖先。

有的对象专门用来存储数据，叫作 POJO 类，还有些用来执行任务，例如，爬虫类或者搜索类。