正则表达式学习笔记

阿左 [1]　　Nobody [2]

December 8, 2012

[1]感谢档
[2]感谢郭嘉

Contents

List of Figures

List of Tables

Abstract

Regex study note

摘要

Regex study note

Part I

基本概念

Chapter 1

简介

如今，人们可以通过电脑来打电话，看电视，给朋友发送即时信息，与其他人玩游戏，甚至可以通过电脑买到你能想到的任何东西，包括从歌曲到 SUV。计算机程序能够通过互联网相互通信使这一切成为了可能。很难统计现在有多少个人电脑接入互联网，但可以肯定，这个数量增长得非常迅速，相信不久就能达到 10 亿。除此之外，新的应用程序每天在互联网上层出不穷。随着日益增加的互联网访问带宽，我们可以预见，互联网将会对人们将来的生活产生长远的影响。

那么程序是如何通过网络进行相互通信的呢? 本书的目的就是通过在 Java 编程语言环境下，带领你进入对这个问题的解答之路。Java 语言从一开始就是为了让人们使用互联网而设计的，它为实现程序的相互通信提供了许多有用的抽象应用程序接口（API，Application Programming Interface），这类应用程序接口被称为套接字（sockets）。

在我们开始探究套接字的细节之前，有必要向读者简单介绍计算机网络和通信协议的整体框架，以使读者能清楚我们的代码将应用的地方。本章的目的不是向读者介绍计算机网络和 TCP/IP 协议是如何工作的（已经有很多相关内容的教程），而是介绍一些基本的概念和术语。

1.1 计算机网络，分组报文和协议

一个程序是作为客户端还是服务器，决定了它在与其对等端（peer）建立通信时使用的套接字 API 的形式（客户端的对等端是服务器，反之亦然）。更进一步来说，客户端与服务器端的区别非常重要，因为客户端首先需要知道服务器的地址和端口号，反之则不需要。如果有必要，服务器可以使用套接字 API，从收到的第一个客户端通信消息中获取其地址信息。这与打电话非常相似：被呼叫者不需要知道拨电话者的电话号码。就像打电话一样，只要通信连接建立成功，服务器和客户端之间就没有区别了。

客户端如何才能找到服务器的地址和端口号呢？通常情况，客户端知道服务器的名字，例如使用 URL (Universal Resource Locator，统一资源定位符) 如 http://www.mkp.com，再通过名字解析服务获取其相应的互联网地址。

获取服务器的端口号则是另一种情况。从原理上来讲，服务器可以使用任何端口号，但客户端必须能够获知这些端口号。在互联网上，一些常用的端口号被约定赋给了某些应用程序。例如，端口号 21 被 FTP (File Transfer Protocol，文件传输协议) 使用。当你运行 FTP 客户端应用程序时，它将默认通过这个端口号连接服务器。互联网的端口号授权机构维护了一个包含所有已约定使用的端口号列表 (见 http://www.iana.org/assignments/port-numbers)。

1.2 什么是套接字

Socket (套接字) 是一种抽象层，应用程序通过它来发送和接收数据，就像应用程序打开一个文件句柄，将数据读写到稳定的存储器上一样。一个 socket 允许应用程序添加到网络中，并与处于同一个网络中的其他应用程序进行通信。一台计算机上的应用程序向 socket 写入的信息能够被另一台计算机上的另一个应用程序读取，反之亦然。

不同类型的 socket 与不同类型的底层协议族以及同一协议族中的不同协议栈相关联，本书只涵盖了 TCP/IP 协议族的内容。现在 TCP/IP 协议族中的主要 socket 类型为流套接字 (sockets sockets) 和数据报套接字 (datagram sockets)。流套接字将 TCP 作为其端对端协议 (底层使用 IP 协议)，提供了一个可信赖的字节流服务。一个 TCP/IP 流套接字代表了 TCP 连接的一端。数据报套接字使用 UDP 协议 (底层同样使用 IP 协议)，提供了一个" 尽力而为" (best-effort) 的数据报服务，应用程序可以通过它发送最长 65500 字节的个人信息。当然，其他协议族也支持流套接字和数据报套接字，但本书只对 TCP 流套接字和 UDP 数据报套接字进行讨论。一个 TCP/IP 套接字由一个互联网地址，一个端对端协议 (TCP 或 UDP 协议) 以及一个端口号唯一确定。随着进一步学习，你将了解到把一个套接字绑定到一个互联网地址上的多种方法。

图 1.2 描述了一个主机中，应用程序、套接字抽象层、协议、端口号之间的逻辑关系。

值得注意的是一个套接字抽象层可以被多个应用程序引用。每个使用了特定套接字的程序都可以通过那个套接字进行通信。前面已提到，每个端口都标识了一台主机上的一个应用程序。实际上，一个端口确定了一台主机上的一个套接字。从图 1.2 中我们可以看到，主机中的多个程序可以同时访问同一个套接字。在实际应用中，访问相同套接字的不同程序通常都属于同一个应用 (例如，Web 服务程序的多个拷贝)，但从理论上讲，它们是可以属于不同应用的。