学习正则表达式 (图灵程序设计丛书)

作者：[美]Michael Fitzgerald

版权信息

书名：学习正则表达式

作者：Michael Fitzgerald

译者：王热宇

ISBN：978-7-115-31149-8

目录

版权声明

O'Reilly Media, Inc.介绍

O'Reilly Media通过图书、在线服务、杂志、调查研究和会议等方式传播创新者的知识。自1978年开始，O'Reilly一直都是发展前沿的见证者和推动者。超级极客正在开创未来，我们关注着真正重要的技术趋势，通过放大那些“微弱的信号”来刺激社会对新科技的采用。作为技术社区中活跃的参与者，O'Reilly的发展充满着对创新的倡导、创造和发扬光大。

作为出版商，O'Reilly为软件开发人员带来革命性的“动物书”，创造了第一个商业网站（GNN），组织开放源代码峰会，以至于开源软件运动以此命名，通过创立Make杂志成为DIY革命的主要先锋，公司一如既往地用各种方式和渠道连接人们和他们所需要的信息。O'Reilly的会议和峰会聚集了超级极客和高瞻远瞩的商业领袖，共同描绘将开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly现在还将先锋专家的知识传递给普通计算机用户。无论是通过印刷书籍、在线服务或者面授课程，每一项O'Reilly的产品都反映了公司不可动摇的信念——信息是激发创新的力量。

业界评论

“O'Reilly Radar博客有口皆碑。”

——Wired

“O'Reilly凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O'Reilly Conference是聚集关键思想领袖的绝对典范。”

——CRN

“一本O'Reilly的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim是位少有的商人，他不光放眼于最长远、最广阔的视野，并且切实地按照Yogi Berra的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去，Tim似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

前言

本书通过示例介绍如何编写正则表达式，旨在让读者轻松掌握正则表达式。事实上，笔者几乎将所涉及的每一个概念都通过示例展示了出来，读者很容易模仿尝试。

正则表达式有助于找到文本字符串中的各种模式。更确切地说，正则表达式是经过专门编写的文本字符串，用来匹配字符串（尤其是文件内字符串）集合中符合该模式的所有字符串。

正则表达式最早出现于美国数学家斯蒂芬·克莱尼编写的*Introduction to Metamathematics*一书中（1952年Van Nostrand公司出版）。但其实这个概念早在20世纪40年代初就已形成。到了70年代，随着Unix操作系统及其实用程序sed、grep等问世，正则表达式得到了计算机科学家更为广泛的使用。Unix操作系统是美国电话电报公司下属贝尔实验室的Brian Kernighan、Dennis Ritchie、Ken Thompson以及其他工作人员的杰作。

据我所知，最早出现正则表达式的计算机应用程序是QED编辑器。QED是Quick Editor的缩写，它是为运行在Scientific Data Systems公司〔1〕SDS 940计算机中的Berkeley Timesharing System编写的。1970年的记录显示，QED是由Ken Thompson在之前MIT的Compatible Time-Sharing System中另外一个编辑器基础上重写而成的。从此，计算技术领域有了真正的正则表达式实现。（附录中的表A-1列出了QED的正则表达式特性。）

〔1〕Scientific Data Systems（英文缩写SDS），是Max Palevsky于1961年在美国成立的一家计算机公司，也是最早在计算机设计中使用集成电路和硅晶体管的公司。SDS计算机主要针对大型科学计算，物美价廉。“太空竞赛”期间NASA曾购买了很多台SDS计算机。SDS在1969年被施乐（Xerox）公司收购，1975年由于管理不善和销售下滑被关闭。在施乐管理期间，该公司一度被称为XDS。——编者注

本书中用来展示示例的工具很多，但多数都容易获取，而且也很实用。只有少数工具目前还没有好用的Windows版本。如果你觉得哪个工具不好用，完全可以不用。但要真正学习正则表达式，我还是建议在Unix环境中学习。我使用Unix环境长达25年，每天仍然能够学到不少新东西。

> “不懂Unix的人注定还要重新发明一个蹩脚的Unix。”——Henry Spencer〔2〕

〔2〕Henry Spencer是加拿大程序员，著名正则表达式库regex的作者。这个正则表达式库被许多程序包或编程语言采用，比如Perl、Tcl和MySQL，等等。在多伦多大学工作期间，Henry Spencer从1981年开始运作美国之外的第一个Usenet站点。这个站点后来被谷歌收购，作为1980年代Usenet的公开档案。另外，他还写过“10 Commandments for C Programmers”（C程序员十诫，http://www.seebs.net/c/10com.html）。——编者注

部分工具可以通过浏览器在线使用，这对于许多读者是十分方便的。其余的工具需要使用命令行和shell脚本，还有一些工具是在桌面上运行的。如果你手头没有这些工具，从网上下载也很方便。其中大多数工具是免费的，偶尔有需要付费的也不贵。

本书中不会出现很多专业术语。我会在必要的时候告诉你正确的术语，但这种情况很少。因为多年的经验表明，专业术语常会造成障碍。换句话说，我会尽可能用通俗易懂的语言描述正则表达式，以免你晕头转向不知所措。因为本书的理念是“略知大概，即可实践”。

正则表达式的实现多种多样。你会发现在vi（vim）、grep及sed等Unix命令行工具中使用的正则表达式也可以在其他程序中见到。各种程序设计语言都支持正则表达式，比如Perl（当然啦〔3〕）、Java、JavaScript、C#、Ruby等。就连XSLT 2.0这样的声明式语言中也有正则表达式。你还会发现Notepad++、Oxygen及TextMate等应用程序同样支持正则表达式。

〔3〕Perl，后来被人们解释为Practical Extraction and Reporting Language的缩写。由这个非官方的“全称”——实用提取和报告语言——可知，Perl在处理文本文件和生成报表方面是非常强大的。1987年，Larry Wall在美国宾夕法尼亚州蓝铃（Blue Bell）地区的Unisys公司当程序员的时候发明了Perl。在该语言后来的发展中，正则表达式功能得到不断丰富和加强，最终成为Perl独树一帜的招牌特色。——编者注

大多数正则表达式实现各有异同。本书不会逐一讨论它们的差异，但也会涉及一些。如果要我把**所有**实现的**全部**不同点都列出来，恐怕非得把我累吐了血不行。所以我在本书中就不纠缠这些细节了。总而言之，如果你期待一本正则表达式的入门书，那就选这本吧。

目标读者

本书适合从零开始学习正则表达式的读者。如果你准备开始学习正则表达式或编写程序，本书就是很好的起点。换句话说，所有听说过正则表达式且对其非常感兴趣，但还没有真正理解正则表达式的人，都是这本书的目标读者。如果你属于这种情况，本书很适合你。

我将采取由简到繁的顺序来介绍正则表达式的特性，也就是先从简单的开始。

如果你已经对正则表达式及其用法有所了解，或者已是位编程老手，本书可能不适合你。本书面向需要指导的初学者。如果你写过一些正则表达式，但希望加深对基本概念的理解，也可以阅读本书，只是我们的节奏可能比你想像的要慢一些。

我在此推荐几本学习完本书后应该阅读的书。第一本是Jeff Friedl的*Mastering Regular Expressions, Third Edition*（参见 http://shop.oreilly.com/product/9781565922570.do）。这本书对正则表达式进行了全面阐述，强烈推荐读者阅读。Jan Goyvaerts和 Steven Levithan执笔的*Regular Expressions Cookbook*（参见 http://shop.oreilly.com/product/9780596520694.do）也是一本相当不错的书[1]。Jan Goyvaerts开发的RegexBuddy是一个强大的桌面应用（参见 http://www.regexbuddy.com），而Steven Levithan开发了我们第1章就会用到的在线正则表达式处理程序RegexPal（参见 http://www.regexpal.com）。

[1] 这两本书的中文版分别是《精通正则表达式（第3版）》（电子工业出版社）和《正则表达式经典实例》（人民邮电出版社）。——编者注

阅读要求

为了更好地学习本书，你需要使用Unix或Linux操作系统上的一些工具。这些工具在Mac上的Darwin系统（BSD在Mac上的衍生系统）、Windows电脑中运行的Cygwin（参见http://www.cygwin.com及http://www.gnu.org）中都能找得到。

本书提供了大量示例供你实验，只是看一遍印象不会深刻。要真正掌握正则表达式，就应该按照这些示例的步骤自己操作，最好把所有示例都过一遍。最好的学习方式是亲身实践，而不是做个旁观者。你得通过本书学会使用高亮显示匹配结果以验证正则表达式的网站、强大的Unix命令行工具、分析正则表达式或用正则表达式搜索文本的桌面程序。

Github上托管着本书的示例代码（https://github.com/michaeljamesfitzgerald/Introducing-Regular-Expressions）。另外，通过http://examples.oreilly.com/0636920012337/examples.zip也可以下载到本书所有的示例和测试文件[1]。学习本书之前，你最好先在自己的计算机上创建一个新文件夹，并将这些文件保存到该文件夹中。

（1）读者也可以从图灵社区本书网页（http://www.ituring.com.cn/book/955）随书下载部分下载本书示例代码。——编者注