sed & awk

第二版

sed与awk

Dale Dougherty & Arnold Robbins 著

张旭东 杨作梅 田丽华 等译

O'REILLY®

机械工业出版社
China Machine Press

HZ BOOKS

sed与awk

第二版

Dale Dougherty & Arnold Robbins 著

张旭东 杨作梅 田丽华 等译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Paris • Sebastopol • Taipei • Tokyo

O'Reilly & Associates, Inc.授权机械工业出版社出版

机械工业出版社

图书在版编目（CIP）数据

sed 与 awk（第二版）/（美）多尔蒂（Dougherty, D.），（美）罗宾斯（Robbins, A.）著；张旭东等译．－北京：机械工业出版社，2003.6
书名原文：sed & awk, Second Edition
ISBN 7-111-11527-9

I. s... II.①多 ... ②罗... ③张... III. UNIX 操作系统 IV. TP316.81
中国版本图书馆 CIP 数据核字（2002）第 008368 号

北京市版权局著作权合同登记
图字：01-202-1827 号

书　　名/ sed 与 awk（第二版）
书　　号/ ISBN 7-111-11527-9
责任编辑/ 贾梅，徐申
封面设计/ Edie Freedman，张健
出版发行/ 机械工业出版社
地　　址/ 北京市西城区百万庄大街 22 号（邮政编码 100037）
经　　销/ 新华书店北京发行所发行
印　　刷/ 北京牛山世兴印刷厂
开　　本/ 787 毫米 × 1092 毫米　16 开本　28 印张　410 千字
版　　次/ 2003 年 6 月第一版　2003 年 6 月第一次印刷
印　　数/ 0001-4000 册
定　　价/ 55.00 元（册）

O'Reilly & Associates公司介绍

为了满足读者对网络和软件技术知识的迫切需求，世界著名计算机图书出版机构O'Reilly & Associates公司授权机械工业出版社，翻译出版一批该公司久负盛名的英文经典技术专著。

O'Reilly & Associates公司是世界上在UNIX、X、Internet和其他开放系统图书领域具有领导地位的出版公司，同时是联机出版的先锋。

从最畅销的《The Whole Internet User's Guide & Catalog》（被纽约公共图书馆评为二十世纪最重要的50本书之一）到GNN（最早的Internet门户和商业网站），再到WebSite（第一个桌面PC的Web服务器软件），O'Reilly & Associates一直处于Internet发展的最前沿。

许多书店的反馈表明，O'Reilly & Associates是最稳定的计算机图书出版商——每一本书都一版再版。与大多数计算机图书出版商相比，O'Reilly & Associates公司具有深厚的计算机专业背景，这使得O'Reilly & Associates形成了一个非常不同于其他出版商的出版方针。O'Reilly & Associates所有的编辑人员以前都是程序员，或者是顶尖级的技术专家。O'Reilly & Associates还有许多固定的作者群体——他们本身是相关领域的技术专家、咨询专家，而现在编写著作，O'Reilly & Associates依靠他们及时地推出图书。因为O'Reilly & Associates紧密地与计算机业界联系着，所以O'Reilly & Associates知道市场上真正需要什么图书。

作者简介

Dale Dougherty 是 Songline Studios 的总裁和首席执行官（CEO），是负责在线内容的O'Reilly & Associates的成员。作为规划出坚果系列的编辑，除了《sed & awk》外，Dale还编写了《DOS Meets UNIX》（与 Tim O'Reilly 合著）、《Using UUCP & Usenet》（与 Grace Todino 合著）和《Guide to the Pick System》。

Arnold Robbins 是亚特兰大人，专业程序员和技术作者。从1980年有人向他介绍在PDP-11上运行的UNIX第6版时，他就开始使用UNIX系统。到1987年开始参与GNU的awk版本——gawk项目时，他已经是awk的重量级用户。作为POSIX 1003.2的支持者，他帮助制订了awk的POSIX标准。他现在是gawk及gawk文档的维护者。这些文档可以从自由软件基金会得到，而且这些已经被SSC编为《Effective AWK Programming》一书出版。

封面介绍

《sed与awk》封面上的动物是瘦小的懒猴。懒猴在夜间活动，生活在树上，是没有尾巴的灵长类动物，有厚的、柔软的毛皮和大而圆的眼睛。主要分布在印度南部和锡兰，在那里它们生活在树上，很少下到地面。可以观察到它们向自己手和足上撒尿——这样做是为了在它们攀登时增加摩擦使它们能紧握树干，并留下气味的轨迹。

这种瘦小的懒猴高度为7到10英寸，重量为12盎司或更少。它依靠吃水果、树叶和捕获小动物为生。

目录

前言

本书介绍了一组名字奇特的UNIX实用工具: **sed**和**awk**。这组实用工具有很多共同的特征，譬如正则表达式在模式匹配中的应用等。模式匹配在sed和awk的使用中是很重要的部分，因此本书详尽地解释了UNIX正则表达式的语法。一般情况下，从**grep**到sed和awk的学习过程是很自然的，所以本书涵盖了上述3个程序，而重点集中在sed和awk。

sed和awk是一般用户、程序员和系统管理员们处理文本文件的有力工具。sed的名字来源于其功能，它是一个字符流编辑器（stream editor），可以很好地完成对多个文件的一系列编辑工作。awk的名字来源于它的开发人Aho、Weinberger和Kernighan，它是一种程序设计语言，非常适合结构化数据的处理和格式化报表的生成。本书强调了awk的POSIX定义。另外，在讨论awk的3个可以免费获得的版本和2个商业版本以前，本书还简要地描述了awk的最初版本，所有这些版本都实现了awk的POSIX定义。

本书的重点是编写sed和awk脚本来快速解决用户各种各样的问题。大多数脚本都可以称为“快速定位”。另外，我们还会涉及到一些需要更仔细地设计和开发，能够解决较大问题的脚本。

本书内容

第一章“强大的编辑工具”，是对sed和awk的特征和功能的概括性描述。

第二章“了解基本操作”，论述了sed和awk的基本操作，并展示了从sed到awk的功能方面的进步。二者共有相似的命令行语法，以脚本的形式接受用户指令。

第三章“了解正则表达式语法”，非常详细地描述了UNIX正则表达式语法。通常，新用户会对这些用于模式匹配的奇怪表达式感到无所适从。掌握正则表达式语法是很重要的，这可以从sed和awk中得到更多的东西。本章中模式匹配的例子主要依赖于**grep**和**egrep**。

第四章“编写sed脚本”，从本章开始，用3章的篇幅对sed进行介绍。本章介绍了那些只使用几个sed命令编写简单的sed脚本的基本要素。还给出了一个可以简化sed脚本调用的shell脚本。

第五章“基本sed命令”和第六章“高级sed命令”，将sed命令分成基本的和高级的命令。基本命令类似于手工编辑命令，而高级命令则介绍简单的编程功能。高级命令包含对保留空间（一个预留的临时缓冲区）的处理命令。

第七章“编写awk脚本”，从本章开始，用5章的篇幅对awk进行介绍。本章介绍了这个脚本化语言的主要特征。介绍了许多脚本，其中包含修改**ls**命令输出结果的脚本。

第八章“条件、循环和数组”，描述了如何使用普通的程序设计结构，例如条件、循环和数组。

第九章“函数”，描述了如何使用awk的内置函数以及如何编写用户定义的函数。

第十章“底部抽屉”，概述了一组不同性质的awk主题。其中包括：如何从awk脚本中执行UNIX命令，如何将输出定向到文件和管道。另外，本章还提供了几个调试awk脚本方面的建议。

第十一章“awk系列产品”，描述了awk最初的V7版本，当前的贝尔实验室的版本，来自自由软件联盟的GNU awk（gawk），以及Michael Brennan编写的mawk等。后面三者都有可以自由获取的源代码。本章还描述了两个商业实现，MKS awk和Thomson Automation awk（**tawk**），以及将类似awk的功能带到Visual Basic环境的VSAwk。

第十二章“综合应用”，给出了两个较长的、更加复杂的awk脚本，它们共同印证了这种语言的几乎所有特征。第一个脚本是交互式拼写检查程序。第二个脚本则处理和格式化一本书的索引或一套书的主索引。

第十三章“脚本的汇总”，给出了用户提供的许多脚本，展示了编写sed和awk脚本的不同的风格和技术。

附录一“sed快速参考”，是描述sed的命令和命令行选项的快速参考。

附录二“awk快速参考”，是awk的命令行选项和它脚本语言完整描述的快速参考。

附录三“第十二章的补充”，给出了第十二章描述的**spellcheck.awk**脚本和**masterindex** shell脚本的完整清单。

sed和awk的实用性

sed和awk是Version 7 UNIX（也称为V7或第七版）的一部分，从那时起它们就成为标准发布的一部分。sed自从被提出以来就没改动过。

自由软件联盟GNU项目的sed版本是可以自由获取的（从技术上讲虽然没有放在公共域中）。GNU sed的源代码可以通过匿名的FTP（注1）从*ftp.gnu.ai.mit.edu*上得到。它存在于文件*/pub/gnu/sed-2.05.tar.gz*中。这是利用**gzip**程序压缩的tar文件，gzip的源代码可以在相同的目录下得到。全球有许多站点对主GNU发布站点的文件做了“镜像”；如果你知道离你最近的站点，就可以从那得到这些文件。注意要使用“binary”或“image”模式传送这些文件。

1985年，awk的作者对awk做了扩充，添加了许多有用的特征。可惜的是，几年以来这个新版本一直只存在于AT&T系统中。从Release 3.1开始它成为UNIX System V的一部分。新的awk名为nawk，旧版本仍然保留原来的名字。System V Release 4系统也是这样。

对商业的UNIX系统（例如来自Hewlett-Packard、Sun、IBM、Digital和其他的系统）来说，命名情况变得更复杂了。所有这些系统都有一些旧的和新的awk版本，但是每个厂商为程序的命名都不同。有的为**oawk**和**awk**，有的为**awk**和**nawk**。我

注1：如果不能访问Internet并且还希望得到GNU sed副本，请联系Free Software Foundation, Inc., 59 Temple Place, Suite 330, Boston, MA 02111-1307 U.S.A。电话号码是1-617-542-5942，传真号是1-617-542-2652。

们能够提供的最好的建议就是检查本地文档（注2）。在本书中，使用术语**awk**来描述POSIX awk，特殊的实现则通过名字来引用，例如“gawk”或“Bell Labs awk”。

第十一章讨论了3个可自由获取的awk（包括从什么地方可以得到它们）以及几个商业版本。

注意：自从本书的第一版以来，awk语言已经被标准化为POSIX Command Language和Utilities Standard（P1003.2）的一部分。所有现代的awk实现都向上与POSIX标准兼容。

P1003.2标准混合了来自新的awk和gawk的特征。在本书中，你可以认为，对POSIX awk的一个实现是对的东西对另一个实现也是对的，除非注明是特别的版本。

DOS版本

gawk、mawk和GNU sed已经移植到DOS系统。在主GNU发布站点上有这些程序的DOS版本的文件。另外，gawk已经移植到OS/2、VMS和Atari与Amiga微型计算机系统中，移植到其他系统（Macintosh、Windows）的工作也正在进行。

egrep、sed和awk可以作为MKS工具包（Mortice Kern Systems, Inc., Ontario, Canada）的一部分用于基于MS-DOS的机器。它们的awk实现支持POSIX awk的特征。

MKS工具包还包括Korn shell，这意味着为UNIX系统上的Bourne shell编写的许多shell脚本都可以在PC上运行。而MKS工具包的大多数用户可能已经发现了这些UNIX中的工具，我们希望这些程序的好处对那些没有大胆地进入UNIX的PC用户来说也是显而易见的。

Thompson Automation Software（注3）有一个用于UNIX、DOS和Microsoft Windows的awk编译器。这个版本很有意思，它拥有用awk编写的awk的一些扩展，还包括一个用awk编写的awk的调试程序。

有时我们也使用PC，因为Ventura Publisher是一个非常大的格式化软件包。我们

注2：纯化论者将新的awk简单地称为“**awk**”，这个新的awk打算取代最初的awk。可是，自发布以来几乎已经10年了，却还是没有取代。

注3：5616 SW Jefferson, Portland, OR 97221 U.S.A.，在美国电话为1-800-944-0139.，在其他地方电话为1-503-224-1639。

喜欢它的原因之一是可以连续使用**vi**创建和编辑文本文件，并使用sed编写用于编辑工作的脚本。我们曾使用sed编写转换程序，从而将**troff**宏转换成Ventura样式表标签。我们还利用它在批处理方式下插入标签。这可以省去必须手工为文件中的重复元素加标签的麻烦。

sed和awk对于编写处理不同的文件格式的转换程序也非常有用。

sed和awk的其他信息源

长时间以来，这些实用工具的主要信息源是包含在*UNIX Programmer's Guide*第2卷中的两篇文章。文章*awk ——A Pattern Scanning and Processing Language*(1978年9月1日）是由awk的3个作者编写的。在这10页中，它提供了一个简要的指南并且讨论了几个设计和实现的问题。文章*SED ——A Non-Interactive Text Editor*（1978年8月15日）由Lee E.McMahon编写。它是一个参考，给出了每个功能的完整描述，并且包含一些很有用的示例（使用Coleridg的*Xanadu*作为示例输入）。

在商业书籍中，sed和awk的最重要的处理出现在由Brian W.Kernighan和Rob Pike合著的《The UNIX Programming Environment》（Prentice-Hall,1984）中。标题为"Filters"的章节不仅解释了这些程序如何工作，而且还展示了它们如何一起来构建有用的应用程序。

awk的作者合著了一本描述其增强版本的书：《The AWK Programming Language》(Addison-Wesley，1988)。它包含许多完整的例子并且论述了可以应用awk的广泛领域。它遵从《UNIX Programming Environment》的风格，有时对于新用户来说太难了。书中示例程序的源代码可以在*netlib.bell-labs.com*的*/netlib/research/awkbookcode*目录下找到。

信息与技术POSIX （Portable Operating System Interface，可移植的操作系统接口）的IEEE标准第2部分：Shell和Utility（标准1003.2-1992）（注4）描述了sed和awk（注5)。这是针对基于sed和awk编写的可移植shell程序所能提供的功能特

注4: 据说那快了3倍！

注5: 这个标准不能在线获取。它可以向IEEE定购，在美国和加拿大的电话为1-800-678-IEEE(4333)，在其他地方的电话为1-908-981-0060。或者在Web浏览器中浏览*http://www.ieee.org/*。费用为U.S. $228，包括标准1003.2d-1994 ——Amendment 1 for Batch Environments。IEEE的会员和/或IEEE协会享受折扣。

征的“官方”描述。因为awk本身就是一种程序设计语言，因此它同样是可移植awk程序的官方描述。

1996年，自由软件联盟出版了由Arnold Robbins编著的《The GNU Awk User's Guide》。这是gawk的文档文件，相比Aho、Kernighan和Weinberger的书，它采用了更多实例教学的方式。它有两个完整的章节完全是示例，并且涵盖了POSIX awk。该书还由SSC以书名《Effective AWK Programming》出版，而且该书的Texinfo源自于gawk的发布。

当前GNU版的sed所存在的最大不足就是缺乏相应的文档，甚至连一页帮助页（manpage）都没有。

在对UNIX的大多数一般性介绍中，对一大串实用工具的介绍时都会提到sed和awk。这些书中，Henry McGilton和Rachel Morgan的《Introducing the UNIX System》提供了基本编辑技巧的最佳处理，包括所有UNIX文本编辑器的使用。

由本书的原作者和Tim O'Reilly合著的《UNIX Text Processing》（Hayden Books，1987）一书完整地概述了sed和awk（虽然没有介绍awk的新版本）。那本书的读者会在本书中发现一些重复的部分，但是从总体上讲这里采用了不同的方法。但在本书中我们将sed和awk区别对待，在假设只有高级用户才会使用awk工具的情况下，这里我们尽量给出与这两者彼此相关的程序。这些不同的工具，可以独立使用也可以相互配合，为文本处理提供令人兴奋的强大功能。

最后，在1995年Usenet新闻组*comp.lang.awk*形成了。如果你在前面提到的书籍中没有找到自己需要了解的知识，你可以在新闻组张贴问题，这是一个可能获得他人帮助的好机会。

这个新闻组会定期张贴一篇“常见问题解答（FAQ）”的文章。除了回答有关awk的问题以外，FAQ还列出了许多站点，从那些站点可以获得用于不同系统的不同awk版本的二进制程序。你可以通过FTP从主机*rtfm.mit.edu*的*/pub/usenet/comp.lang.awk/faq*文件中检索到FAQ。

示例程序

本书中的示例程序最初是在运行A/UX 2.0（UNIX System V Release 2）的Mac IIci和运行SunOS 4.0的SparcStation 1上编写和测试的。要求POSIX awk的程序使用gawk 3.0.0和来自Bell Labs FTP站点的Bell Labs awk的August 1994版本进

行了重新测试（参看第十一章有关FTP的详细内容）。sed程序用SunOS 4.1.3 sed和GNU sed 2.05进行了重新测试。

获取示例源代码

可以通过从O'Reilly & Associates的Internet服务器上获得本书中程序的源代码。本书的示例程序可以用多种电子方式获得：FTP、Ftpmail、BITFTP和UUCP。最先列出的是最便宜、最快速和最容易的方式。如果你从上至下读取，第一个为你工作的可能是最好的。如果你直接和Internet相连就使用FTP。如果你没有连接到Internet上，但是可以向Internet站点（包括CompuServe用户）发送和接收电子邮件，那就使用Ftpmail。如果你能够通过BITNET发送电子邮件则使用BITFTP。如果上面的方式都不能工作就使用UUCP。

FTP

为了使用FTP，需要一台可以直接访问Internet的机器。以下是一段示例，黑体字是你应该键入的：

```
$ ftp ftp.oreilly.com
Connected to ftp.oreilly.com.
220 FTP server (Version 6.21 Tue Mar 10 22:09:55 EST 1992)ready.
Name (ftp.oreilly.com:yourname):anonymous
331 Guest login ok,send domain style e-mail address as password.
Password:yourname@domain.name (在此使用你的用户名和主机名)
230 Guest login ok,access restrictions apply.
ftp> cd /published/oreilly/nutshell/sedawk_2
250 CWD command successful.
ftp> binary (很重要，必须为压缩文件指定二进制传送方式)
200 Type set to I.
ftp> get progs.tar.gz
200 PORT command successful.
150 Opening BINARY mode data connection for progs.tar.gz.
226 Transfer complete.
ftp> quit
221 Goodbye.
```

这个文件是gzip压缩的tar档案文件；通过键入下面的语句从档案文件中提取这些文件：

```
$ gzcat progs.tar.gz | tar xvf -
```