国外计算机科学教材系列

数据压缩原理与应用

（第二版）

Data Compression

The Complete Reference, Second Edition

[美] David Salomon 著

吴乐南 等译

電子工業出版社
Publishing House of Electronics Industry
http://www.phei.com.cn

国外计算机科学教材系列

数据压缩原理与应用

(第二版)

Data Compression

The Complete Reference, Second Edition

[美] David Salomon 著

吴乐南 等译

電子工業出版社

Publishing House of Electronics Industry

北京·BEIJING

内容简介

数据压缩是现代计算最重要的领域和工具之一。从获取数据到CD-ROM，从编码理论到图像处理，现代计算的许多层面都依赖于数据压缩。本书对数据压缩的许多不同类型和方法提供了全面的参考。内容包括详尽而有益的分类、最常用方法的描述、方法使用和获益的比较以及“如何”应用的讨论。全书的介绍沿数据压缩领域的主干来组织、游程编码、统计方法、基于字典的方法、图像压缩、音频压缩和视频压缩。该书的主要主题为：视频压缩、小波方法、音频压缩、用于JPEG和JBIG的QM编码器、图像变换、用于压缩简单图像的EIDAC方法、前缀图像压缩、ACB和FHM曲线压缩、几何压缩和边缘破碎法。

本书为所有的计算机科学家、计算机工程师、电气工程师、信号/图像处理工程师，以及其他需要一部压缩方法大全的科学家们，提供了一本十分宝贵的参考和指南。

版权贸易合同登记号：图字：01-2002-1394

图书在版编目（CIP）数据

数据压缩原理与应用=Data Compression: The Complete Reference：第2版/（美）萨洛蒙（Salomon. D.）著；吴乐南等译. -北京：电子工业出版社，2003.9
（国外计算机科学教材系列）
ISBN 7-5053-8247-0

Ⅰ.数… Ⅱ.①萨… ②吴… Ⅲ.数据压缩-教材 Ⅳ. TP274

中国版本图书馆CIP数据核字（2003）第076657号

责任编辑：谭海平
印 刷 者：北京兴华印刷厂
出版发行：电子工业出版社 http://www.phei.com.cn
北京市海淀区万寿路173信箱 邮编：100036
经 销：各地新华书店
开 本：787 × 1092 1/16 印张：38.5 字数：985千字
版 次：2003年9月第1版 2003年9月第1次印刷
定 价：59.00元

译 者 序

约从1990年开始加速的一场“数字化革命”，给人类社会带来了深刻而长远的影响。在当今社会，人们已深知“信息”之重要：“食指”上网，“拇指”发“信”（全国人民仅拜年就发了几十亿条“短信”！）；同样，人们也乐于享受“数码”的优越性：VCD/DVD光盘、GSM/CDMA手机、电视机顶盒、计算机多媒体、数码照相机/录像机等等，美不胜收。所有这些，都需要压缩数字编码的数据或“数码”，也都得益于本书之主题——数据压缩，即用最少的数码来表示信号。说得更完整些，就是以最少的数码表示信源所发的信号，减少容纳给定消息集合或数据采样集合的信号空间。

因此，数据压缩已成为当今数字通信、数字广播、数字存储、数字摄影、数字出版和多媒体娱乐中的一项关键性的共性技术，甚至还有人杜撰出“Compressionism”（压缩主义）一词来表达这种共识。数据压缩的“经典”原理只是信息论中的信源编码理论，但由于数据的类型不同，用户的要求各异，市场的空间巨大，并且立即就能“见效”，因此，这一领域一直在增长，而且是高速增长。一方面，人们的胃口越来越大，要“数字化人体”，“数字化地球”，数字化一切希望用计算机处理、存储、展现和交互的对象；另一方面，各种不同的理论、技术、算法甚至数学分支被市场导向到这一领域竞争、比试，“胜者为王”，成为国际认可的标准，败亦欣然，说不定几年后又“东山再起”。因此，数据压缩方法繁多且更新快，现有书籍没几年就可能过时。相当多的读者可能需要有一本工具书或手册，能够介绍更多的数据压缩原理和收录更多的数据压缩算法，供其筛选、比较和研究。显然，重在讲述基本原理的教科书和拿来就能用的程序集，都难以满足这一要求。而David Salomon先生所著的这本书，恰好填补了这一空白。纵览全书，译者至少感觉到了“全、新、浅、趣”四大特点。

首先是“全”。正如著者本人所言：“为所有的计算机科学家、计算机工程师、电气工程师、信号/图像处理工程师，以及其他需要一部压缩方法大全的科学家们，提供了一本十分宝贵的参考和指南”，“向非专业读者清晰地阐述数据压缩的原理和现用的所有重要方法”。只要看一下本书的目录，就完全可以同意这种观点。其中有许多内容（特别是第8章中的一些方法）在其他书籍中很难找到。

其次是“新”。本书第二版比第一版又新增了第5章、第6章和第7章整整三章，以及其他大量的内容，对于JPEG-LS、JBIG-2和JPEG 2000等新标准都有介绍。当然，现在成为应用主流且需求最甚的对于视频和音频的压缩，第6章、第7章两章的更新还不够快，有关H.263、MPEG-4和H.264的内容尚来不及反映，相信会在第三版中出现。但是，瑕不掩瑜，本书在文本和静止图像等方面的内容，应该是既新又全、极其出色的。

第三是“浅”。一方面是指没有冗长的数学推导，另一方面则是指对读者的先修课程要求不多，“必需的数学背景已减至最少，只限于对数、矩阵、多项式、微积分，以及概率的概念”，这就便于广大科技人员甚至业余爱好者自学。著者称本书所打算面向的读者群，是那些“具有计算机科学的基础知识，懂一点编程和数据结构，对诸如比特、兆、ASCII、文件、I/O以及二进制查找等术语感到亲切，希望懂得数据压缩原理”的人们。这当然也包括高等院校中那些跨专业选修课程的本科和专科学生。但是，著者又表明：“本书不打算指导软件实现，因此没有什么程

序”。这可能又使某些希望“拿来就用”的软件公司开发人员有些失望。然而，不能说没有程序就不通俗、不实用，译者以为，有时读很多注释不够且来路不明又调试不通的“程序”，反而更加“痛苦”：“折腾”一番之后，简直怅然若失！

最后是“趣”。即“趣味性”，这又体现在几方面。一是材料“有趣”，这不仅体现在主要内容上，而且组织在背景材料中。譬如，“Lena”是图像处理界和图像压缩中广泛使用的一幅测试图像，多少莘莘学子用它发表了论文。可译者从不知道也从未想过它（“她”）的出处，直到看了本书（4.5节），才了解到原来还有这么一段“有趣”的故事。二是能吸引读者的“兴趣”，著者的指导思想是“尽量少用数学，多给出些例子、图表和习题”。而据译者统计，仅第8章正文中的图表即多达544幅！平均每页都有图或表，堪称图文并茂。另外，书中的例子和习题都很有特色，特别是习题与课文互为补充，而几乎所有习题的答案都在网上，可以“偷看”，不至于因冥思苦想不得其解而兴趣索然。三是著者的“风趣”，译者对此感触尤深。特别是著者文学功底深厚，广证博引，从西方的《圣经》，到东方的《道德经》；从文学名著，到插科打诨，把一部专业性很强的工具书，组织得这样“有趣”，能吸引译者（相信还会有相当多的读者）的“兴趣”，又体现了著者的“风趣”，真乃大家风范。但也正是这些“只言片语”，在让译者忍俊不禁、拍案叫绝之余，也颇令译者汗颜：要把这些在特定的上下文环境下“只可意会，不可言传”的神来之笔，“信、达、雅”地翻译给读者，真正是勉为其难，经常于似懂非懂之中“推敲”多时，仍难免以“江郎才尽”而告终。

本书是由东南大学无线电工程系专门从事数据压缩的有关师生共同翻译的，是集体劳动的成果。参加本书初译的是：赵正敏副教授（第1章）、曹振新博士（第2章）、王淑兰讲师（第3章）、吴冬升博士（第4章）、陈小蓄博士（第5章）、李中科博士（第6章）、谭亮硕士（第7章）和司宏伟硕士（第8章）；王太君副教授复译了第6章、第7章；吴乐南教授翻译了全书其他部分和图表，并对全书内容进行了复译、审校和统稿。

书中部分错误已按著者放在网上的勘误表加以更正。新发现的个别疏漏也以“译注”的形式标出。但由于才识所限，又时间匆匆，疏漏之处仍然不少，还祈望读者不吝赐教！

吴乐南

2003年于南京

第二版序

第二版的面世有三个原因。首先是应许多热心读者的建议，比如有人这么说：

我刚看完您有关数据压缩的书，太有意思了。在一册仅约20毫米厚的书中能有这么多的算法，它本身就是数据压缩的一个例子！

——Fred Veldmeijer, 1998

原因之二是作者和读者在第一版中发现的错误。它们被列于本书的网址中（见后面），在第二版中得到了更正。

原因之三是本书的书名（最初是由出版商选定的）。为了使书的名字名副其实，必须使本书成为完整的参考书。因此，本书第二版中又增加了许多压缩方法和大量的背景资料。最重要的增添和改动如下：

- 新增了3章。首先是第5章，有关小波及其在图像和声音压缩方面应用的主题相对较新（不大为人所知）。本章首先用连续小波变换（CWT）直观地解释小波，然后详细举例说明如何用Haar变换压缩图像；接着是有关滤波器组和离散小波变换（DWT）的一般性讨论，并且列出了许多常用小波滤波器的小波系数；最后，描述了一些使用或基于小波的重要压缩方法，包括Laplacian金字塔、集分割等级树（SPIHT）、利用零树的嵌入编码（EZW）、指纹压缩的WSQ方法和JPEG 2000（一种用于压缩静止图像的有前途的新方法，见5.19节）。
- 其次新增的是第6章，讨论视频压缩。本章首先概述有关CRT的操作，以及基本的模拟和数字视频的概念，然后继续讨论视频压缩，最后描述MPEG-1和H.261。
- 第三个新增的是第7章，主题是音频压缩。本章首先介绍人类听觉系统的特性，以及如何利用它来得到有损的音频压缩；然后讨论几种简单的声音压缩方法，最后描述MPEG-1的3个音频层，包括非常流行的mp3格式。

其他新材料的组成如下：

- 条件图像RLE（1.4.2节）。
- 标量量化（1.6节）。
- JPEG、JPEG 2000及JBIG所用的QM编码器在2.16节。
- 2.19节讨论上下文树加权，而4.24节则将其扩展用于无损图像压缩。
- 3.4节讨论一种叫做重复次数的滑动缓冲存储器方法。
- 3.25节还包括了有关专利的麻烦事项。
- 相对不太熟悉的Gray码在4.2.1节讨论，与图像压缩有联系。
- 4.3节讨论图像压缩的直觉方法，例如亚采样和矢量量化。
- 4.4节讨论图像变换的重要概念。详细描述离散余弦变换（DCT），介绍了Karhunen-Loève变换、Walsh-Hadamard变换和Haar变换。而4.4.5节则暂时离题讨论离散正弦变换，它与DCT很相近，但性能欠佳，鲜为人知。
- JPEG-LS是无损和近无损图像压缩的国际新标准，在4.7节中介绍。
- JBIG2是另一个国际新标准，用于二值图像压缩，现在可在4.10节找到。
- 4.11节讨论压缩简单图像的EIDAC法，其主要创新是用了两部分上下文。像素P的

面内上下文由其位平面上的若干邻近像素组成;而 P 的面间上下文由趋于与其相关的像素组成,尽管它们位于不同的位平面上。

- 新的 4.12 节涉及矢量量化,后面几节则是自适应矢量量化和块截断编码(BTC)。
- 块匹配是 LZ77(滑窗)适应图像压缩的一种方法,可在 4.14 节找到。
- 差分脉冲编码调制(DPCM)现在包括在新的 4.23 节中。
- 块分解法是一种压缩离散色调图像的有趣方法,见 4.25 节。
- 4.26 节讨论二进制树预测编码(BTPC)。
- 前缀图像压缩与四叉树有关,是 4.27 节的主题。
- 与四叉树有关的另一种图像压缩是四截面。4.28 节将其连同有关的二截面、八截面一起讨论。
- 第一版中有关 WFA 的 4.31 节有误,在 Karel Culik 和 Raghavendra Udupa 的大力帮助下已彻底重写。
- 胞元编码在 4.33 节讨论。
- DjVu 是一种不寻常的方法,打算用于扫描文件的压缩,由贝尔实验室(Lucent Technologies)开发,5.17 节讨论。
- 5.19 节讨论关于静止图像压缩的新的 JPEG 2000 标准。
- 8.4 节是基于排序的上下文相似性方法。此方法按与 ACB 类似的方式利用一个符号的上下文。它还把秩分给符号,这一特性使它与 Burrows-Wheeler 方法及符号秩建立了联系。
- 稀疏串的前缀压缩法添加在 8.5 节。
- FHM 是压缩曲线的非传统方法,使用 Fibonacci 数、Huffman 编码和 Markov 链,8.9 节讨论。
- 8.10 节的跟随特别适合于半结构文本的压缩,它基于无上下文语法。
- 8.11 节详细描述边缘破碎机,这是一种压缩三角形网格连接信息的原始方法。该法及其不同的扩展方法可能成为多边形表面压缩方法的标准,而多边形表面是计算机图形学中最常用的表面类型。边缘破碎机是几何压缩法的一个例子。
- 考虑到空间的限制,所有附录均已删除,它们可以在本书的网址上以 PDF 格式免费下载。附录包括:(1)ASCII 码(包括控制字符);(2)空间填充曲线;(3)数据结构(包括哈希);(4)纠错码;(5)有限状态自动机(多种压缩方法需要它,如 WFA、IFS 和动态 Markov 编码);(6)概率元素;(7)插值多项式。
- 删去了习题答案,它们可以在本书的网址上找到。

目前,本书的网址是作者个人网址的一部分,位于 http://www.ecs.csun.edu/~dsalomon/。域名 BooksByDavidSalomon.com 已经保留,并将始终指向任何未来的网址。作者的电子信箱为 david.salomon@csun.edu,但是已计划把任何发给 < 任何名称 > @BooksByDavidSalomon.com 的电子邮件转发给作者。

读者也可以从 http://welcome.to/data.compression 重定向到本书的网址。发到 data.compression@welcome.to 的电子邮件也将被重定向。

那些对数据压缩感兴趣的读者通常应翻阅一下书末题为"加入数据压缩协会"的短节,以及两个 URL 站点:http://www.internz.com/compression-pointers.html 和 http://www.hn.is.uec.ac.jp/~arimura/compression_links.html。

David Salomon

第一版序

历史上，数据压缩并不属于计算机科学的首要领域之一。该领域的工作者似乎先要在第一个 20～25 年的时间来发展足够的数据，才能感觉到需要压缩。今天，当计算机领域约有 50 年历史的时候，数据压缩成了一个宽广而活跃的领域，还有巨大的市场。对于这一点，也许最好的证明就是数据压缩会议（Data Compression Conference，DCC）的名声。

用于压缩不同类型数据的原理、技术和算法正在被许多人快速地开发出来，其基础概念借鉴了来自统计学、有限-状态自动机、空间-填充曲线以及傅里叶和其他变换等相当广泛的理论。这一趋势自然引发了关于此主题的许多书的出版，这就提出了一个问题：为什么还要出一本关于数据压缩的书？

答案显然是因为这个领域很大，而且一直在增长，因此“创造”出更多的潜在读者，并使得现有的书籍没几年就过时了。

写作本书的最初理由，是向非专业读者清晰地阐述数据压缩的原理和现用的所有重要方法。作者的本意是，描述和讨论能为有一些计算机使用和操作背景的读者所理解的内容。因此，尽量少用数学，多给出些例子、图表和习题。书中并不严格证明每条断言，而是多次代之以“能够显示……”或“能够证明……”的说法。

习题是本书尤为重要的特色，与课文互为补充，任何有兴趣充分理解数据压缩和书中所述方法的人都应该去做这些习题。作者提供了几乎所有习题的答案（在本书的网页上），但是在查看答案之前，读者显然应尽力做出每道习题。

致谢

我特别想感谢 Nelson Beebe，他非常仔细地润色了第一版的全部文本，做了许多校正和建议。也十分感谢如下人员：Christopher M. Brislawn，复审了 5.18 节并允许我们使用图 5.64；Karel Culik 和 Raghavendra Udupa，对加权有限自动机（WFA）给出了实质性的帮助；Jeffrey Gilbert，润色了 4.25 节（块分解）；John A. Robinson，审查了 4.26 节（二进制树预测编码）；Øyvind Strømme，审查了 5.10 节；Frans Willems 和 Tjalling J. Tjalkins，审查了 2.19 节（无上下文加权）；Hidetoshi Yokoo，对于 3.15 节和 8.4 节的帮助。

作者还想感谢 Paul Amer，Guy Blelloch，Mark Doyle，Hans Hagen，Emilio Millan，Haruhiko Okumura 和 Vijayakumaran Saravanan，他们帮助找出了错误。

我们似乎对于物体的收缩和扩张有着本能的迷恋。因为我们在这方面的实际本领很有限，所以我们爱读有关人和物体能戏剧性地改变其自然形状的小说，例如 Jonathan Swift 的“*Gulliver's Travels*”（1726），Lewis Carroll 的“*Alice in Wonderland*”（1865）和 Isaac Asimov 的“*Fantastic Voyage*”（1966）。

“*Fantastic Voyage*”刚开始是著名作家 Isaac Asimov 所写的一部电影剧本，而当该影片正在拍摄的时候（该片 1966 年发行），Asimov 又把它重写为一部小说，改正了电影剧本中某些最突出的缺陷。其情节讲述了在一艘潜艇上把一组医学科学家收缩到微观大小后，注入一位病人

的体内，以便利用激光束除去病人脑部的血块。关键在于那位病人 Benes 博士是一位科学家，他改进了这一小型化过程，并最先使其实用化。

由于电影和小说都获得了成功，Asimov 后来又写了一部“*Fantastic Voyage* Ⅱ：*Destination Brain*”，但这部小说成为一大败笔。

David Salomon

目　录

引　言

Giambattista dell Porta，一名文艺复兴时期的科学家，也是1558年出版的“*Magia Naturalis*”一书的作者，书中他讨论了很多主题，包括鬼神学、磁学和暗箱。这本书提到一个后来被称为“共振电报”(sympathetic telegraph)的假想装置，该装置由两个类似于指南针的圆盒组成，各有一根磁针。每个盒子在通常标记方位之处都代之以26个字母，而其要点是两根针应由同一块磁石磁化。Porta假设这样可使磁针协调工作，当一个盒子里的磁针指向一个字母时，另一个盒子里的磁针也将转向同一个字母。

不用说，这样一个装置无法运转(这毕竟比Samuel Morse早了近300年)，但是在1711年，一位焦急的妻子写信给伦敦的“*Spectator*”期刊，询问如何忍受亲爱的丈夫长期不在身边。忠告人Joseph Addison提供了几条可行的建议，然后提到了Porta的装置，说一对这种盒子可以使她和丈夫保持联络，即使他们“被密探和守卫监视，或者被城堡和艰险阻隔”。然后Addison先生补充道，当被情侣们使用时，共振电报的面板上除了26个字母外，还应包括“一些在情书中经常出现的完整单词”。比如“I love you”这条消息，此时就只需发送3个符号，而不是10个。

这一建议就是本文压缩的一个早期例子，压缩的取得是给常用消息编短码而给其他消息编长码。甚至更重要的是，这展现了数据压缩的概念是如何自然地来到了对通信交流感兴趣的人们面前的。我们好像生来就知道：为了节省时间，发送的数据越少越好。

数据压缩是把输入数据流(源流或原始数据)转变为另一种较小的数据流(输出流或压缩流)的过程。流即存储器中的一份文件或一块缓存。数据压缩之所以流行，有两个原因：(1)人们喜欢积攒数据而不愿丢弃任何东西，不论有多大的存储设备，迟早都会溢出。数据压缩看来是有用的，因为它能延缓这一不可避免的进程；(2)人们讨厌长时间地等待数据的传输，当我们坐在电脑前等待网页打开或文件下载时，自然会觉得多于几秒钟都好像是太长的一段等待时间。

数据压缩是在过去20年成熟起来的，这一领域文献的数量和质量充分证明了这一点。其实，人们早就感到要压缩数据，甚至是在电脑出现以前。

数据压缩有很多有名的方法。它们基于不同的理念，适合不同的数据类型，产生不同的效果，但是原理都相同，即通过去除源文件的原始数据中的冗余度来压缩数据。任何非随机选择的数据都有一定的结构，可利用这种结构得到数据的更小的表示(看不出什么结构性)。术语冗余度(redundancy)和结构(structure)用于专业文献，此外还有平滑(smoothness)、相干(coherence)及相关(correlation)，都是指同一个意思。因此，在有关数据压缩的任何讨论中，冗余度是一个相当重要的概念。

例如在典型的英文文本中，字母E最常出现而Z很少出现(表1和表2)，称为字母冗余(alphabetic redundancy)。这启示我们为字母安排变长码，使E的码字最短而Z的最长。另一类冗余度是文本冗余(contextual redundancy)，通常表现为字母Q后常跟有字母U(即在简单英语中，某些双字母及三字母比其他模式更常见)。图像冗余度则表现为，在一幅非随机的图像中，相邻像素的颜色往往相近。

2.1节讨论信息论并给出一个冗余度的定义。然而，即使我们不知道该名词的精确定义，

直觉上也清楚变长码比定长码的冗余度小(甚至没有冗余度)。定长码便于文本处理,所以是一种有用的但却有冗余的码。

通过减少冗余度来压缩的思想暗示了数据压缩的通用法则,即"对常见事件(符号或短语)赋短码,对稀有事件赋长码"。实现这一法则的方法很多,一项对任何压缩方法的分析表明,它们在本质上都遵循这一通用法则。

数据压缩的实现是把低效(长的)表达方式改为高效(短的)表达方式,因此有可能压缩只是因为通常计算机中数据的格式比绝对需要长。采用低效(长的)数据表示的原因是使数据易于处理,而数据处理要比数据压缩更普遍和更重要。有一个很好的例子:字符的 ASCII 码表示要长于它的绝对需要,用 7 位代码是因为定长码易于处理。然而变长码更有效,因为有些字符比其他字符更常用,因此可以赋予更短的码字。

在一个总是用最短的可能格式表示数据的世界中,无法压缩数据。此时,作者们将不写关于数据压缩的书,而是写关于如何为不同类型的数据确定其最短格式的书。

点拨一下聪明人……

数据压缩领域的主要目标,当然是开发好之又好的压缩方法。然而数据压缩艺术的主要困惑之一是何时不再追寻更好的压缩。经验表明有限地调节算法来从数据中榨取出所剩下的最后几位冗余度,回报递减。修改算法使压缩改进 10%,则可能导致运行时间增加 10%,而程序复杂度增加得还要高。Fiala 和 Greene 想出了一种好方法来解决这一矛盾(3.7 节)。在开发了主要算法 A1 和 A2 后,他们又将其修改为速度更快但压缩性能稍差的算法 B1 和 B2。然后他们反过来修改 A1 和 A2,牺牲速度但得到了稍好的压缩。

去除冗余度的压缩原则也回答了下述问题:"为什么已压缩的文件不能进一步压缩了?"回答当然是这样的:文件已经没有或只有很少的冗余度,所以没有什么可以去除了。例如,随机文本就是这类文件,每个字母等概率出现,因此对其都赋以定长码并不增加冗余度。当我们压缩这样的文件时,没有冗余度可去除(另一种回答是,如果可以进一步压缩已压缩的文件,则相继的压缩都可以减小文件尺寸,直到它只剩下一个字节甚至一位。这显然是荒谬的,因为一个孤立字节不可能代表任意大小的文件所包含的信息)。读者可以参看一下 8.7 节关于压缩随机数据的一个有意思的歪曲。

既然提到了随机数据,那我们就多说两句。通常,极少有随机数据的文件,但有一个很好的例子——已压缩的文件。压缩文件的所有者一般都知道该文件已压缩过了,不会尝试进一步压缩,但有一个例外——通过调制解调器传数据。现代调制解调器都有自动压缩所传数据的硬件,如果数据已压缩,就不会有进一步的压缩,甚至还可能扩展。这就是为什么调制解调器应"在线"监测压缩比。如果压缩比太低,就应停止压缩,把其余数据不压缩地传出去。V.42 协议(3.18 节)就是该技术的一个好例子。2.7 节将讨论压缩随机数据的"技巧"。

习题 1:(有趣)找出含有 5 个元音"aeiou"并按此原始顺序出现的英文单词。

数据压缩已变得如此重要,以致一些研究人员(例如,参见[Wolff 99])提出了 SP 理论[代表简洁(simplicity)和能力(power)],认为一切运算都是压缩!具体地说,它将数据压缩理解为去除信息中不必要的复杂度(冗余度)的过程,从而在保留尽可能多的无冗余描述能力的同时,得到最大程度的简洁。SP 理论基于以下推测:

- 把各类计算和形式推理视为用模式匹配、联合及搜索来进行信息压缩有助于理解。

- 找到并去除冗余度的过程总可以在基础上理解为搜索互相匹配的模式,把任何模式的重复实例合并或联合成一个。

本书讨论许多压缩方法,一些适用于文本数据,另一些则适用于图形数据(静止图像或电影)。大多数方法可分为以下4类:游程编码(RLE)、统计方法、基于字典的(有时称为LZ)方法和变换。第1章和第8章讨论基于其他原理的方法。

在深入展开之前,我们先讨论一下数据压缩的重要名词。

- 压缩器(compressor)或编码器(encoder)是压缩输入流中的原始数据,并建立由已压缩的(低冗余度)数据构成的输出流的程序。解压缩器(decompressor)或解码器(decoder)则进行相反的转换。注意,术语"编码"(encoding)很常用,有很广泛的含义,但既然我们只讨论数据压缩,编码器就是指数据压缩器。术语编解码器(codec)有时用于同时指编码器和解码器。类似地,术语压扩(companding)是"压缩/扩展"的简写。
- 本书一直用术语"流"来代替"文件"。流更常用,因为压缩数据可直接传给解码器,无需写成文件再保存。同样,待压缩的数据可能是从网上下载来的,而不是从文件输入的。
- 对于原始的输入流,我们称之为未编码(unencoded)、原(raw)或原始(original)数据。最终压缩流中的内容被称为已编码(encoded)或已压缩(compressed)数据。术语"位流"(bitstream,或比特流)也常用于专业文献,指的就是已压缩的流。

表1 英文字母出现的概率

字母	频率	概率	字母	频率	概率
A	51 060	0.072 1	E	86 744	0.122 4
B	17 023	0.024 0	T	64 364	0.090 8
C	27 937	0.039 4	I	55 187	0.077 9
D	26 336	0.037 2	S	51 576	0.072 8
E	86 744	0.122 4	A	51 060	0.072 1
F	19 302	0.027 2	O	48 277	0.068 1
G	12 640	0.017 8	N	45 212	0.063 8
H	31 853	0.044 9	R	45 204	0.063 8
I	55 187	0.077 9	H	31 853	0.044 9
J	923	0.001 3	L	30 201	0.042 6
K	3812	0.005 4	C	27 937	0.039 4
L	30 201	0.042 6	D	26 336	0.037 2
M	20 002	0.028 2	P	20 572	0.029 0
N	45 212	0.063 8	M	20 002	0.028 2
O	48 277	0.068 1	F	19 302	0.027 2
P	20 572	0.029 0	B	17 023	0.024 0
Q	1611	0.002 3	U	16 687	0.023 5
R	45 204	0.063 8	G	12 640	0.017 8
S	51 576	0.072 8	W	9244	0.013 0
T	64 364	0.090 8	Y	8953	0.012 6
U	16 687	0.023 5	V	6640	0.009 4
V	6640	0.009 4	X	5465	0.007 7
W	9244	0.013 0	K	3812	0.005 4
X	5465	0.007 7	Z	1847	0.002 6
Y	8953	0.012 6	Q	1611	0.002 3
Z	1847	0.002 6	J	923	0.001 3

本书预印版中26个字母出现的频率和概率，总共有708 672个字母(大写和小写)，组成约145 000个单词。

大多数但不是所有的专家认为英语中出现频率最高的字母依次为ETAOINSHRDLU(经常被写为两个孤立的单词ETAOIN SHRDLU)。不过，[Fang 66]中有另一种观点。最常用的双字母(2字母组合)为TH、TA、RE、IA、AK、EJ、EK、ER、GJ、AD、YU、RX和KT。而绝大多数单词以S、P、C开始，用E、Y、S结尾。

表2 字符的出现频率和概率

字符	频率	概率	字符	频率	概率	字符	频率	概率
e	85 537	0.099 293	x	5238	0.006 080	F	1192	0.001 384
t	60 636	0.070 387	|	4328	0.005 024	H	993	0.001 153
i	53 012	0.061 537	-	4029	0.004 677	B	974	0.001 131
s	49 705	0.057 698	)	3936	0.004 569	W	971	0.001 127
a	49 008	0.056 889	(	3894	0.004 520	+	923	0.001 071
o	47 874	0.055 573	T	3728	0.004 328	!	895	0.001 039
n	44 527	0.051 688	k	3637	0.004 222	#	856	0.000 994
r	44 387	0.051 525	3	2907	0.003 374	D	836	0.000 970
h	30 860	0.035 823	4	2582	0.002 997	R	817	0.000 948
l	28 710	0.033 327	5	2501	0.002 903	M	805	0.000 934
c	26 041	0.030 229	6	2190	0.002 542	;	761	0.000 883
d	25 500	0.029 601	I	2175	0.002 525	/	698	0.000 810
m	19 197	0.022 284	^	2143	0.002 488	N	685	0.000 795
\	19 140	0.022 218	:	2132	0.002 475	G	566	0.000 657
p	19 055	0.022 119	A	2052	0.002 382	j	508	0.000 590
f	18 110	0.021 022	9	1953	0.002 267	@	460	0.000 534
u	16 463	0.019 111	[	1921	0.002 230	Z	417	0.000 484
b	16 049	0.018 630	C	1896	0.002 201	J	415	0.000 482
.	12 864	0.014 933	]	1881	0.002 183	O	403	0.000 468
1	12 335	0.014 319	'	1876	0.002 178	V	261	0.000 303
g	12 074	0.014 016	S	1871	0.002 172	X	227	0.000 264
0	10 866	0.012613	_	1808	0.002 099	U	224	0.000 260
,	9919	0.011 514	7	1780	0.002 066	?	177	0.000 205
&	8969	0.010 411	8	1717	0.001 993	K	175	0.000 203
y	8796	0.010 211	‘	1577	0.001 831	%	160	0.000 186
w	8273	0.009 603	=	1566	0.001 818	Y	157	0.000 182
$	7659	0.008 891	P	1517	0.001 761	Q	141	0.000 164
}	6676	0.007 750	L	1491	0.001 731	>	137	0.000 159
{	6676	0.007 750	q	1470	0.001 706	*	120	0.000 139
v	6379	0.007 405	z	1430	0.001 660	<	99	0.000 115
2	5671	0.006 583	E	1207	0.001 401	"	8	0.000 009

本书预印版中93个字符出现的频率和概率，共有861 462个字符。

金色甲虫

到此为止，我们已开始入门了。上表的一般用途很明显，但若用这种特殊代号，对我们没多大帮助。因为我们最主要的字符是8，先假定它是自然的字母“e”。为检验这种假定，我们来观察一下8是否经常成对出现(因为“e”在英文中成对出现的频率很高)在这些词中，如“meet”，“fleet”，“speed”，“seen”，“been”，“agree”，等等。我们看到它在本例中成对出现不少于5次，尽管这种密码很简洁。

——Edger Allan Poe