ELSEVIER

Broadview
www.broadview.com.cn

深入搜索引擎

—— 海量信息的压缩、索引和查询

Managing Gigabytes:
Compressing and Indexing Documents and Images Second Edition

[新] Ian H.Witten [澳] Alistair Moffat [新] Timothy C.Bell 著
梁斌 译

Google发源地
斯坦福大学
信息检索课程
首选教材之一

深入、细致地讲述搜索引擎是怎样运行的……

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

深入搜索引擎

——海量信息的压缩、索引和查询

[Managing Gigabytes:
Compressing and Indexing Documents and Images Second Edition]

[新]Ian H.Witten [澳]Alistair Moffat [新]Timothy C.Bell 著
梁斌 译

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

内 容 简 介

本书是斯坦福大学信息检索和挖掘课程的首选教材之一，并已成为全球主要大学信息检索的主要教材。本书理论和实践并重，深入浅出地给出了海量信息数据处理的整套解决方案，包括压缩、索引和查询的方方面面。其最大的特色在于不仅仅满足信息检索理论学习的需要，更重要的是给出了实践中可能面对的各种问题及其解决方法。

本书作为斯坦福大学信息检索课程的教材之一，具有一定的阅读难度，主要面向信息检索专业高年级本科生和研究生、搜索引擎业界的专业技术人员和从事海量数据处理相关专业的技术人员。

Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition
Ian H. Witten, Alistair Moffat, Timothy C. Bell
ISBN-13:978-1-55860-570-1

版权贸易合同登记号 图字：01-2003-2038

图书在版编目（CIP）数据

深入搜索引擎：海量信息的压缩、索引和查询 /（新）威顿（Witten,I.H.），（澳）莫夫特（Moffat,A.），（新）贝尔（Bell,T.C.）著；梁斌译.—北京：电子工业出版社，2009.6
书名原文:Managing Gigabytes：Compressing and Indexing Documents and Images, Second Edition
ISBN 978-7-121-08491-1

I. 深… II.①威…②莫…③贝…④梁… III.互联网络－情报检索－高等学校－教材 IV.G354.4

中国版本图书馆 CIP 数据核字（2009）第 035102 号

责任编辑：孙学瑛
印　　刷：北京智力达印刷有限公司
装　　订：北京中新伟业印刷有限公司
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：35.5　字数：574 千字
印　　次：2009 年 6 月第 1 次印刷
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。

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。

服务热线：（010）88258888。

About the Chinese edition: a word from the authors

Ni hao! It is a great pleasure to have the opportunity to write a few words for this new Chinese edition of *Managing Gigabytes*. We produced the first edition of this book in 1993—when the World-Wide Web was hardly known. With the advent of the web the techniques we described became far more important than before. Indeed, we have been told that our book was required reading at Google in the early days. Back in 1993 a gigabyte was still considered to be an enormous amount of storage space, but it turns out that the techniques described apply equally well to terabytes—even petabytes—of data. Of course, the book has been considerably updated since that first edition.

We have all visited China and seen many wonders: ranging from Jade Dragon Snow Mountain in the west to Shanghai in the east; from Beijing and Yinchuan in the north to Nanjing and Guangzhou in the south. The first trip by one of us was over thirty years ago, and we have seen the country blossom in technical maturity and sophistication from a base that can only be described as *mamahuhu*. We are absolutely delighted that our book will contribute to further extended growth in the key technologies of full-text searching and data compression.

Many years ago one of our students showed us a Chinese version of an early edition of *Managing Gigabytes*. The translation, which we had not been aware of, was embarrassingly poor, and a devastatingly prominent error was that only one author's name appeared on the cover of the book—indeed the other two were not mentioned anywhere inside either (although their families were faithfully thanked in the translated preface).

So you can imagine how extraordinarily pleased we are to see this new translation of the second edition of our book. During visits to China in 2008 and 2009 Tim Bell taught the material in this book at Huazhong University of Science and Technology, and could

see the value of having a translation available for students. We were delighted when we heard from 梁斌, who was starting on a translation around that time, and even more pleased when we realised what a careful and thorough job he was doing.

Perhaps the greatest compliment anyone can pay an author is to painstakingly translate their work. We are deeply and sincerely grateful to 梁斌 for the tremendous amount of time and energy he has put into this project. *Xiexie*! We know that he has mastered the material because he has pointed out a number of obscure errors in the English edition, and has asked us many penetrating questions.

We hope you will benefit from his effort, and enjoy reading this book.

Timothy C. Bell

Alistair Moffat

Ian H. Witten

March 2009

原著赞誉

Witten，Moffat 和 Bell 的第二版中不仅仅有更新、更好的文本搜索算法，而且还有大量有关图像分析和图像文本处理的知识。如果你关心搜索引擎，你就会需要这本书，这是目前唯一能够细致入微到搜索引擎如何运作的各个细节的一本书籍。这本书不仅翔实而且可读性强，作者将顶尖的程序和完美的写作风格融为一炉。

——Michael Lesk，国家自然基金会

对每个希望掌握大规模数据处理的从业人员来说，这本书是一本圣经。在 Infoseek 公司，我们要求每个搜索工程师阅读此书。作者的这项工作令人赞叹，他们已经把近 5 年内信息检索研究界最令人瞩目的成果写进了本书。

——Steve kirsch，Infoseek 公司创始人

能够包括压缩、文件组织、全文索引技术和文档管理系统，因此本书无疑是无以伦比的。学生，研究者和从业人员将会从本书中受益

——Bruce Croft,马萨诸塞大学智能信息检索中心主任

快速响应和高效存储时超媒体研究者和开发者的基础技术，我强烈向大家推荐这本可读性强且发人深思的好书。

——Rob Aksycn, Knowledge Systems 公司

译者序

1998 年从美国斯坦福大学产生了一段传奇的财富神话，这就是今天市值约千亿美元的 Google。众所周知，Google 正是由 Lawrence Page 在斯坦福大学发起的研究项目转变而来的。正是由于斯坦福大学对全球信息检索的杰出贡献，译者从事相关研究的时候也曾阅读了大量出自斯坦福大学的课件、论文和推荐教材。

在这些资源[1]中，《Managing Gigabytes》，简记做“MG”，是其中一本极其重要的书籍。在译者集中学习信息检索的 2005 年，这本书是斯坦福大学信息检索和挖掘课程[2]的首选教材之一，和 MIR[3]一起成为全球主要大学信息检索的主要教材。

MG 深入浅出地给出了海量信息数据处理的整套解决方案，包括压缩、索引和查询的方方面面。本书理论性较强，公式众多，很多数据的给出并没有做具体的解释，此外还包括一些文化背景差异带来的理解障碍。但是作者和译者联手为大家奉献了 412 个注解，协助大家更好地理解本书。

和 MIR 不同的是，MG 更加具有实践性，这得益于 3 位作者精心编写的 MG 检索引擎，该检索引擎被实践证明具有很强的易用性和伸缩性，附录 B 介绍的新西兰电子图书馆就使用了 MG 代码作为其内核。MG 源代码可以在原著的官网上找到。本书绝大部分算法和思想都在代码中被完整体现，是不可多得的学习和实践材料。

---

[1] Information Retrieval Resources:http://www-csli.stanford.edu/~hinrich/information-retrieval.html.

[2] Text Retrieval and Mining: http://www.stanford.edu/class/cs276a/syllabus2004.html.

[3] MIR = Modern Information Retrieval, by R. Baeza-Yates and B. Ribeiro-Neto.

本书主要面向信息检索专业方向的研究生、从事搜索引擎相关工作和其他对搜索技术感兴趣的人们，除了从书中获取严谨的理论知识以外，还可在 MG 源代码上展开实际的研究。无论从哪一点来看，本书都是非常好的研究起点。

本书作者 Ian H.Witten，Alistair Moffat 和 Timothy C.Bell 均是信息检索领域赫赫有名的专家，特别是 Timothy C.Bell 教授在本书的翻译过程中给予了巨大的帮助，同时译者也为原著的勘误做出了贡献[4]。

本书的第一版曾由科学出版社于 1996 年翻译并出版，由于时代局限、技术落后等原因，其内容和原著有一定距离。译者在学习和翻译的过程中都获得了它的帮助，这里对参与本书第一版翻译工作的同志表示感谢。

最后要特别感谢包括原著 3 位作者在内的信息检索专家们无私地分享了他们的技术成果，并且感谢电子工业出版社大力引进，编辑孙学瑛女士及方方面面工作人员给予的帮助。由于译者能力有限，若有翻译不当之处，欢迎发送电子邮件至 mgigabyte@gmail.com 批评指正。

最后引用本书中的一段原话作为结尾：“在信息科学技术的历史上，从来没有像今天这样，创造如此大的价值的如此多的技术却掌握在如此少的人的手里。”希望能够和原著的作者一样做出自己一份微薄的贡献。

梁斌

2009 年 4 月 15 日

[4] MG 在线勘误：http://www.cs.mu.oz.au/mg/errata.html。

前言

计算机革命使得全社会都再也不能离开信息。然而，大部分的信息还是以其原始的格式存储着，即数据(data)。原始信息是海量的。这些数据主要产生于商业活动、法律诉讼和政府活动。随之而来的还有不计其数的复制品，这主要是报道、杂志和报纸产生的。最后这一切存储在档案馆、图书馆和计算机中。面对的挑战是如何高效和有效地处理大量的信息，以便能方便、廉价地定位和抽取有效的信息。

从空间的角度看，在纸张上存储文档的传统方法是昂贵的，更重要的是，当需要定位和检索所需要的信息时，需要付出高昂的代价。因此能够经济地存储和访问文档就变得越来越重要。几百英尺高的一大堆书中所包含的文本只需要一块磁盘就可以存下，从物理空间占用的角度看，电子媒体的这种存储能力是惊人的。和人工的文档索引方法相比，这种方法即具有伸缩性（全部的单词都可以作为关键词）和可靠性（因为索引构造的过程完全不需要人的参与，也就没有人为干扰）。此外，当今社会的各类组织不得不处理各种来源的电子信息，例如，机器可读文本、传真、其他扫描文档和数字图像。和纸媒体相比，使用电子媒体在存储和访问上都特别有效。

这本书讨论如何管理大量文档，GB 的数据。1GB 大约是 1000MB，这足够存储 1000 本书籍，相当于从地板摞到天花板这么高的书籍。在日常生活的词汇不断地增长的同时，大规模存储设备容量也在不断增长。就在 20 年前，百兆数据的需求看上去是那么的奢侈，甚至是幻想。今天个人电脑已经配置上了 GB 的存储设备，甚至一些小的机构也需要存储数 GB 的数据。自从本书第一版问世以来，万维网爆炸般地创造了万亿字节（terabytes）的公开数据，让越来越多的人意识到处理如此大规模数据的难题是特别重要的。

管理如此大量数据主要需要面对两个挑战，这两个挑战都在本书中进行了讨论。第一个挑战是如何有效地存储数据。这主要通过压缩的方法来实现。第二个挑战是提供一种通过关键词搜索的方法来提供快速访问信息的方法。因此，一个特别定制

的索引尤为关键。传统的压缩和搜索方法需要调整以适应这些挑战。这也是本书中主要讨论的两个主题。本书讨论的这些技术应用的结果是确保计算机系统可以存储数百万的文档和能够在秒级的时间内检索到包含任给关键词（或关键词组合）的文档，甚至可以在不到 1s 内完成查询。

举个例子来说明本书中所讨论的这些方法的威力。掌握了这些方法后，你可以对数 GB 的文本创建一个数据库，并且使用它来响应类似这样的查询请求，“在仅适用工作站的条件下，用数秒时间就能在全部文档中检索同时包含‘managing’和‘gigabytes’的段落”。事实上，如果能够对文本创建恰当的索引，这并不是什么神奇的事情。最令人着迷的是这些创建的数据库（包括索引和完整文本），当然都是压缩过的，只有不到原文本的一半大小。不仅如此，创建这样一个数据库只需要数小时即可。最令人惊讶的可能是如果数据集不压缩的话，查询时间还会更少。

大部分本书讨论的技术都已经被提出、实验和应用到实践中。为快速搜索和检索而构造的文本索引被仔细地检查过，这些信息构成了本书的核心。话题还包括文本压缩和建模，压缩图像的方法，压缩文本图像（例如扫描或传真文档）和为了区分图片图表和文本而进行的页面布局识别等。

全文索引不可避免会非常巨大，因此制作的成本也很高。然而，本书揭示了全部单词，如果需要的话，还包括全部数字建立完整索引的奥秘，并阐述了如何用如此小的存储代价支持如此快速的访问能力的技巧。

本书的目标是介绍管理大量文档和图片数据集的最新方法。在阅读本书以后，你将掌握这些技术并同时对它们的威力产生敬意。

随书软件

一个阐述本书思想的完整的系统，MG（代表“Managing Gigabytes”），已经被开发出来。MG 完整代码可以在互联网上自由获得（官方首页 www.cs.mu.oz.au/mg/）。代码用 ANSI C 语言编写并且能够运行在 UNIX 操作系统下，这是一个我们开发的可操作的技术样例。它用一种完整的方法压缩、存储和访问了文本集合、扫描文档和图像。任何布尔型的关键词组合都可以用在对全部文档进行的检索中，同时支持非常规的排名查询（用户仅仅指定一个关键词列表，系统能够让被检索出的相关文档有序排列）。考虑到早先提到的查询例子，在全部文档中检索同时包含‘managing’和‘gigabytes’的段落。在包含 750 000 个文档的数据库中，相当于 2GB 的文本，

对于 MG 来说只需要 1s 就能够访问和解码这两个单词的索引项，这两个单词分别出现了 159458 和 961 次，同时包含这两个单词的文档有 554 个，大约 7MB。取出和解压这些文档只需要不到 1min。

读者定位

对本书感兴趣的读者包括这样几类。对这些主题有兴趣的一般读者；需要掌握信息管理新技术的信息专家。愿意了解技术细节的其他读者；阅读此书的读者包括：信息系统的实践者、程序员、顾问、图书管理员、信息传播者、教授、学生、开发人员、需求工程师、专利检查员，以及对新技术感到好奇的人们。需要发布 CD-ROM 数据库（例如书籍，大百科全书，甚至计算机软件）的人员将直接从本书所阐述的技术中获益。为了避免要求读者具备较多的专业理论和数学知识，除了那些比较难懂的书中在右侧空白处用浅灰色条块标记的部分[1]，读者可以跳过这些部分，并不会影响阅读的连续性。我们对主要的结论均在文中显著给出。

本书可以用于高年级大学生、研究生和专业人员的基础课学习。每一章都介绍了全文检索系统的不同部分，包括文本、索引和图片的压缩方法；大部分的章节可以独立作为短期课程的教材。例如，第 2 章是一个文本压缩方法的完整综述，可以用来作为关于压缩的一个短期课程教材。事实上可以用一本书的篇幅来写这一部分，事实上，他们也这么做了（和 John G.Cleary、和本书的两位作者一起合作了一本叫做 Text Compression 的书）。这个章节提供了一个独立成篇，对实践中常用的方法给出了一个实际的指南，给与那些愿意在这个领域从事工作的人们提供了足够的参考信息。类似的，第 6 章也是独立成篇的，介绍了图像压缩的当前技术和国际标准。第 5 章包括了使用布尔查询和排名查询的信息检索基本概念，给出了关于如何实现的一些具体技术细节。

这本书的组织让两组章节提供深入和更细的子领域的技术细节。第 1，3，4 和 5 章用做研究生关于信息检索的基础课。而第 6，7 和 8 章构成了有关图像分析和压缩的

[1] 译者注：本书翻译版没有用浅灰色条块标记，但书中对可以跳过的地方均显著说明，这些部分相对理论性更强，跳过这些部分不会妨碍理解本书主要内容。

独立模块。更完整的高年级本科生和研究生的关于信息系统和数据压缩的课程所涉及的全部内容都可以在本书中找到，或者作为信息系统和实践数据结构的补充教材。

最后，如果你只对概念感兴趣，对技术细节不感兴趣的话，可以阅读本书第 1 章和最后一章以了解一般的信息。第 1 章介绍了需要解决的问题和给读者一个现实世界的例子，交代了制作一个词汇索引在过去是多么耗时，以及后来他们是怎么被全文检索系统取代的过程。本书需要传达的主要思想：压缩和索引大规模文本和图像集合的方法。第 10 章展望了未来的发展和这些新技术的应用场合。其中一个开发方向是将广播和多媒体信息集成到索引的检索系统中来。这种需求是显然的；任何可以被关键词检索的信息类型都可以整合到压缩的索引系统中来，任何压缩对信息压缩的方法也都可以被引入。将来这类系统将会迅速应用于存储各种大量信息的场合中。

更新和修订的内容

本书的第一版于 1994 年出版，1999 年 3 月，我们出版了它的第二版。在这 5 年间，信息世界中发生了巨大的变化，万维网的繁荣，数字图书馆的创意，信息国际化，Java 语言和网络计算机，卧室中的虚拟现实。今天，最大的信息系统是随处可见的 TV、杂志和广告。今天信息工作者经历了这种冲击和每天都不可避免的大规模数据检索需求所导致的沮丧。这些都在这 5 年内发生了。其中本书中包含的诸多深奥话题中有关文本图像压缩的内容已经成为国际标准，并且很快就能应用到你的传真机上。然而 1993 年预言的一些变化还没有发生：例如，第二版没有被叫做 Managing Terabytes，在第一版中我曾这样预言过。有关技术预言的内容就是这么多。

一方面，全世界的信息已经融化进我们日常的生活中，这在某种程度上延续了我们在 1993 年的预言。另一方面，本书的话题并没有过时：事实上，这些内容和目前的现实更加契合。压缩和索引文档和图像的需求更加强烈。压缩、信息科学和全文检索的基本想法，包括图像表示的基本想法都是相同的。压缩的全文索引的想法特别不寻常。就目前我们了解的情况，非商业的搜索引擎已经基本使用了我们所提到的这些技术：他们付出了巨大的努力，使用了巨大的磁盘和安装了许多内存。他们不存储文本，只存储索引。在出现技术错误时，已经从“Bus error:core dumpled”这样奇怪的提示改为“404 Not Found: The requested URL was not found on this server”，这看上去更加友好。和第一本书出版的时候一样，现在正当时。

虽然第二版的基本核心内容和第一版相同，但是我们尽最大努力更新了部分内

容以反应这五年来发生的变化。当然，我们改正了一些错误，这些错误来自于从在线勘误的积累。事实上，发现的错误出乎预料地少，我们希望第二版错误会更少。第二版的在线勘误可以在 www.cs.mu.oz.au/mg/中找到。我们仔细地编辑了各个章节并且使这些内容保持与时俱进，追加了一些信息参考内容到“进一步阅读”中。最让人感兴趣的部分都追加了新内容，这些就是其中主要的追加。

第 2 章追加了关于文本压缩的最近发展，包括块排序方法（Burrows-Wheeler 转换），近似算术编码和快速哈夫曼编码算法。有些方法的一些细节也进行了追加，效果比较也更新到了最近压缩程序的水平，相对结果采用了最新的 Canterbury 语料，而不是此前使用的 Calgary 语料。

第 3 章讨论了索引技术，追加了关于基于上下文索引压缩的一节内容，包括对最新发展起来的插值编码进行了讨论。关于签名文件和其与倒排文件的效果比较的内容都进行了进一步的修订。

第 4 章中追加了 4 节。第一个新增节讨论了分块倒排索引的使用，这能够支持快速布尔查询。第二个新增节讨论了基于频率排序的倒排索引，这能够改善排名查询的效果。第三个新增节阐述了一些关于运作万维网搜索引擎的一些话题。第四个新增节分析了分布式查询。这些节介绍了排名查询，TREC 项目被进一步修订以包含这 5 年来的一些进展。

第 6 章包括了一些关于无损图像压缩的新内容，包括目前广泛使用的图像事实标准（GIF 和 PNG），一个叫做 CALIC 的高性能无损图像压缩算法。和 JPEG 无损和 JPEG-LS 草案都已经申请成为新的无损压缩的国际标准。

第 7 章追加了关于 JBIG2 一节，这是一个即将成为文档图像压缩的国际标准。虽然直到 2000 年末该方案还没有最终确定，但是在本书出版之时，这个方案有可能会确定下来，其中会包含本书中介绍的许多技术。

第 9 章修订和更新了许多实验结果以反映压缩技术的当前水平和计算机硬件在这 5 年间所取得的成就。特别地，特别细致的一节（包含若干新图例）被追加用来阐述限长的哈夫曼编码。

第 10 章包含了关于因特网和万维网的一些新内容，关于数字图书馆的一些新内容，关于 Web 搜索引擎的新内容和基于代理的信息检索。

书中的附录 B 是一个应用本书思想的一个大型应用，新西兰数字图书馆。这是

一个互联网上可以访问的公开信息资源，使用 MG 作为其内核。它试图展示 MG 软件在信息检索方法的易用性和灵活性。附录解释了提供的一些功能和机制。

最后，从我们在教学中使用该书所获得的经验来讲，教材提供了一个“指导附件”包含了关于本书的问题复习、实验和使用中遇到的问题。这是一本单独的小册子，教师们可以向 Tim C.Bell 索要[2]。地址：Department of Computer Science, University of Canterbury, Christchurch, New Zealand。也可以发邮件到 jsj@phei.com.cn 索要。

致谢

写致谢总是一件令人愉快的事情，许许多多的人都帮助过我们，更高兴有机会能向他们表示感谢。许多在数据压缩和信息检索领域享有盛名的同事在本书的编写过程中给予了大量的鼓励和帮助，尤其是 Abe Bookstein，Nigel Horspool，Tomi Klein，Glen Langdon，Timo Raita 和 Jim Storer。从他们身上我们学到了许多东西，并把其中一部分体现在本书中。特别需要指出的是 John Cleary，RadFord Neal，Ron Sacks-Davis 和 Justin Zobel，我们长期在一起努力工作，这些成绩也是他们的。Bob Kurse 在本书一个重要问题上给出了很有价值的建议，Rod Harries 和 Todd Bender 为词汇索引提供了有用的信息和建议。在这 5 年间，还有其他几位也为我们提供了直接或间接的帮助：Gill Bryant，Sally Jo Cunningham，Tony Dale，Daryl D'Souza，Mike Fleischmann，peter Gutmann,Jan Pedersen，Bill Pennebarker，Art Pollard，Marcelo Weinberger 和 Ross Wilkinson。David Abrahamson 在本书编写工作的早期做了大量工作，他帮我们确定了书中应该包含哪些内容和不该包含哪些内容。他还鼓励我们进行文本图像压缩工作，并且为第 7 章提供了一些素材。我们还要感谢那些评论家，他们最先说服出版社支持我们。当手稿即将完成时，他们又一次给与我们帮助，，如 Ron Murray，Rob Akscyn，Robert Gray，David Hawking，PaulKantor，Yann LeCun，Michael Lesk，Darryl Lovato，Karen SparckJones，Jan Pedersen 和 Peter Willett。Douglas Campbell 对第二版提供了特别细致和有价值的评价。

Morgan Kaufmann 出版社的 Jenifer Mann 和 Karyn Johnson 对第二版的排版工作

[2] 译者注：指导附件也可以与译者联系索要，并请证明您是一位用于教学目的的大学教师，译者会给您免费邮寄复印件。译者电子邮件：mgigabyte@gmail.com。

付出了巨大的努力，Elisabeth Beller 是本书的产品编辑，在整个过程中为我们提供了理想的服务。来自 IBM 的 Joan Mitchell 在本书的编写、修改到出版过程中均给与了许多有价值的帮助。

我们的许多学生也给予了极大的帮助。加拿大 Calgary 大学的 Mary-Ellen Harrison 和 Mark James 以及新西兰 Canterbury 大学的 Hugh Emberson 在我们研究文本图像压缩的过程中给予了极大的帮助。新西兰 Waikato 大学的 Stuart Inglis 和 Abdul Saheed 完成了文档布局识别和采用基于模型压缩技术实现的模式匹配。Caig Nevill Manning 参与我们早期关于索引压缩技术的研究并给予了很多实际的帮助。澳大利亚墨尔本大学的 Peter Thompson 为第 5 章提供了素材。我们还要对那些参与本书许多实验的同学们表示感谢，他们是 Tim A .H.Bell（来自墨尔本大学，请不要与本书的作者 Tim C.Bell 混淆），Gwenda Bensemann，Mike Ciaarella，Craig Farrow，Andrew Kelly，Alex McCooke，Chris Stephens，John Tham，Bert Thompson，Lang Stuiver，Andrew Turpin 和 Glenn Wightwick，他们一起努力的工作聚沙成塔为本书提供了许多宝贵的意见。

在完成第二版的同时，我们要特别感谢 Auckland 大学的 Peter Fenwick，他协助提供了关于 Burrows-Wheeler 转换的素材。Nasir Memon 友好提供了第 6 章的一些信息，JPEG-LS 中大量的信息来源于他写的一篇论文。PaulHoward 对尚处襁褓中的文本图像压缩新标准的描述进行了审阅。Harold Thimbleby 对附录提供了有价值的评价，Yvonne Simmons 对索引部分提供了帮助。Andrew Turpin 对 huffword 程序提供了改进的实现方法，以及关于第 9 章中限长编码的结果提供了帮助。Owen de Kretser 和 Lang Stuiver 对第 3 章和第 4 章中许多结果进行了重新计算。Tim A.H. Bell 对本书的全篇进行了事先地校对工作，Tetra Lindarto，Elizabeth Ng 和 Bronwyn Webster 对校对工作也提供了帮助。Nelson Beebe 就是有价值信息的一个来源，从本书第一版出版之日起就不断给我们鼓励。

第一个 MG 系统是在澳大利亚研究学会和联合信息技术研究院的支持下开发成功的，得到了 Lachlan Andrew，Gray Eddy 和 Neil Sharman 的帮助。从那以后，又有许多人参与进来：Owen de Kretser，Tim Shimming 和 William Weber，以及对该项目有直接贡献的人们。还有许许多多在各个方面做出贡献的人们，由于篇幅所限不能一一列举他们。最小完美哈希函数子程序由昆士兰大学的 Bohdan Majewski 所写，非常感谢他提供在本书中使用这些程序，还有许许多多的人提供了超出我们预期的关于技术细节方面的巨大帮助。 第 6 章的全部图都由 Neil Sharman 制作，包括第 2 章和附录 A 的若干图。第 7 章和第 8 章的很多图例由 Kerry Guise 和 Stuart Inglis 制作。

目录

第 1 章　概览 1

1.1　文档数据库（document databases） 7
1.2　压缩（compression） 10
1.3　索引（indexes） 12
1.4　文档索引 16
1.5　MG 海量文档管理系统 20
1.6　进一步阅读 21

第 2 章　文本压缩 23

2.1　模型 26
2.2　自适应模型 29
2.3　哈夫曼编码 32
范式哈夫曼编码 38
计算哈夫曼编码长度 44
总结 51
2.4　算术编码 51
算术编码是如何工作的 53
实现算术编码 56
保存累积计数 59
2.5　符号模型 61
部分匹配预测 61
块排序压缩 64
动态马尔科夫压缩 69
基于单字的压缩 71

2.6 字典模型.....73
自适应字典编码器的 LZ77 系列.....74
LZ77 的 Gzip 变体.....78
自适应字典编码器的 LZ78 系列.....79
LZ78 的 LZW 变体.....81
2.7 同步.....84
创造同步点.....84
自同步编码.....87
2.8 性能比较.....89
压缩性能.....91
压缩速度.....94
其他性能方面的考虑.....97
2.9 进一步阅读.....98
第 3 章 索引.....102
3.1 样本文档集合.....106
3.2 倒排文件索引.....110
3.3 压缩倒排文件.....115
无参模型（Nonparameterized models）.....117
全局贝努里模型.....120
全局观测频率模型（Global observed frequency model）.....123
局部贝努里模型（Local Bernoulli model）.....124
有偏贝努里模型（Skewed Bernoulli model）.....125
局部双曲模型（Local hyperbolic model）.....127
局部观测频率模型（Local observed frequency model）.....128
上下文相关压缩（Context-sensitive compression）.....130
3.4 索引压缩方法的效果.....132
3.5 签名文件和位图.....134
签名文件.....135
位片签名文件（Bitsliced signature files）.....139
签名文件分析.....144
位图.....147
签名文件和位图的压缩.....148