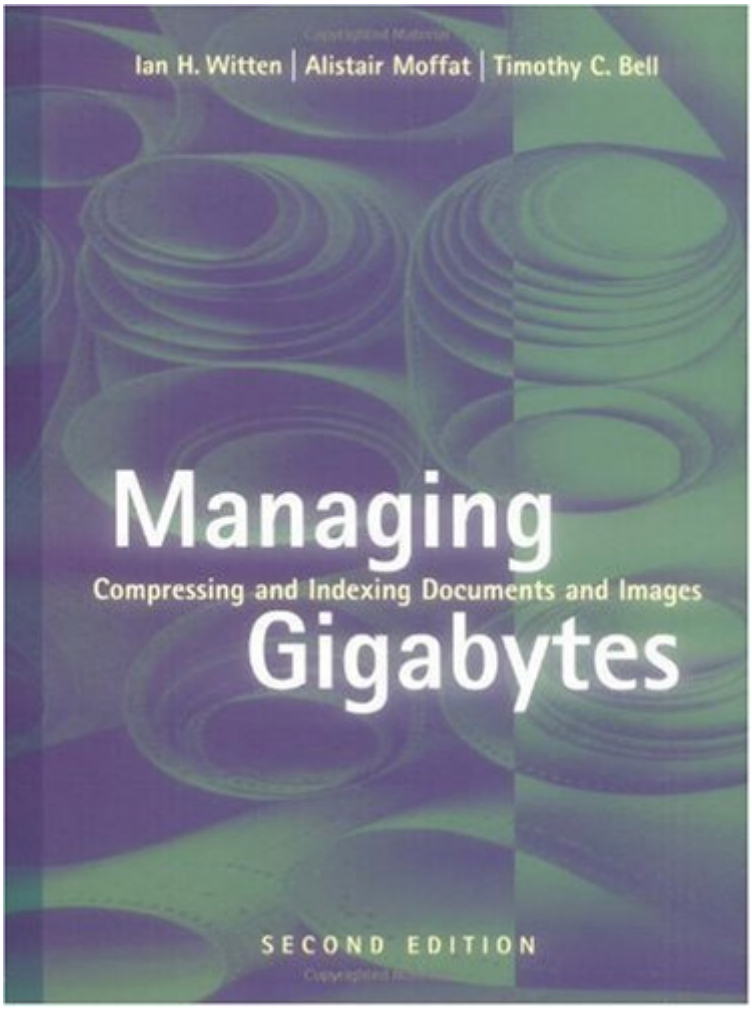
Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition (The Morgan Kaufmann Series in Multimedia Information and Systems)

Ian H. Witten, Alistair Moffat, Timothy C. Bell
*ebooks | Download PDF | *ePub | DOC | audiobook*



#1204770 in eBooks 1999-05-11 1999-05-11File Name: B00440E0PS | File size: 50.Mb

Ian H. Witten, Alistair Moffat, Timothy C. Bell : Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition (The Morgan Kaufmann Series in Multimedia Information and Systems) before purchasing it in order to gage whether or not it would be worth my time, and all praised Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition (The Morgan Kaufmann Series in Multimedia Information and Systems):

70 of 72 people found the following review helpful. The Wonderful Thing Is: It's the Only OneBy Peter NorvigThis is the only book there is that will actually teach you how to build an information retrieval system (aka search engine). It

discusses all the algorithms and tradeoffs, and comes with free downloadable source code to experiment with. Some of the material is standard, but covered in more implementation detail here than anywhere else. Some of the material is novel: you won't find better coverage of compression unless you hand-assemble twenty research papers, and reverse-engineer them to figure out how they're implemented. But with "Managing Gigabytes", it's all here. (Although, after a particularly envigorating discussion of how to string together a bunch of techniques to compress their corpus and save a couple 100MB, I did a check and found you could buy 512MB of RAM for less than the cost of the book. Knowledge is Power, but sometimes a little cash is more powerful.) The only negative is that this book is not called "Managing Terabytes", as the first edition promised/threatened it might be. RAM and disk are cheap, but not that cheap, and for now terabytes (and sometimes petabytes) are managed only by NASA, Google, and a few others. I can't wait to see the third edition!6 of 6 people found the following review helpful. Great Book on Information RetrievalBy A CustomerManaging Gigabytes is the best book out there on information retrieval. If you're interested in implementing your own IR system, there's nothing available that comes close to this book. But the book is good not just because it's the only one out there: the writing is excellent, the algorithms are presented clearly and explained well, and the coverage is thorough. Additionally, the coverage of compression algorithms is the best I've found in any book. All algorithms and pseudo-code in the book are presented clearly enough such that any competent programmer should be able to implement them. If all else fails, however, the free downloadable source code for the mg system can fill in any gaps.All in all, this is the best computer science book I've purchased in years. I wish all CS books were written like this one: it doesn't skimp on the theory or on the implementation details.11 of 11 people found the following review helpful. Best text available. Has no competition.By A CustomerThis text sets the standard for future information retrieval texts and has replaced the Salton books as the canonical academic text.The second edition is highly readable and contains a thorough updating of the algorithms and data structures in the field.I like the text because of its readability, conciseness, thoroughness, and attention to detail. The comparisons of algorithms on realistic sized collections is unparalleled in other texts.I have used this text for the past 5 years in a graduate level information storage and retrieval class but I believe it has a much wider audience due to the quality of writing.Additionally, the free availability of the mg system which implements many of the best algorithms of the text allows the reader/student to take advantage of the technology without having to start from scratch.Highly recommended.

In this fully updated second edition of the highly acclaimed Managing Gigabytes, authors Witten, Moffat, and Bell continue to provide unparalleled coverage of state-of-the-art techniques for compressing and indexing data. Whatever your field, if you work with large quantities of information, this book is essential reading--an authoritative theoretical resource and a practical guide to meeting the toughest storage and access challenges. It covers the latest developments in compression and indexing and their application on the Web and in digital libraries. It also details dozens of powerful techniques supported by mg, the authors' own system for compressing, storing, and retrieving text, images, and textual images. mg's source code is freely available on the Web.Up-to-date coverage of new text compression algorithms such as block sorting, approximate arithmetic coding, and fat Huffman codingNew sections on content-based index compression and distributed querying, with 2 new data structures for fast indexingNew coverage of image coding, including descriptions of de facto standards in use on the Web (GIF and PNG), information on CALIC, the new proposed JPEG Lossless standard, and JBIG2New information on the Internet and WWW, digital libraries, web search engines, and agent-based retrievalAccompanied by a public domain system called MG which is a fully worked-out operational example of the advanced techniques developed and explained in the bookNew appendix on an existing digital library system that uses the MG software

.com Of all the tasks programmers are asked to perform, storing, compressing, and retrieving information are some of the most challenging--and critical to many applications. Managing Gigabytes: Compressing and Indexing Documents and Images is a treasure trove of theory, practical illustration, and general discussion in this fascinating technical subject. Ian Witten, Alistair Moffat, and Timothy Bell have updated their original work with this even more impressive second edition. This version adds recent techniques such as block-sorting, new indexing techniques, new lossless compression strategies, and many other elements to the mix. In short, this work is a comprehensive summary of text and image compression, indexing, and querying techniques. The history of relevant algorithm development is woven well with a practical discussion of challenges, pitfalls, and specific solutions. This title is a textbook-style exposition on the topic, with its information organized very clearly into topics such as compression, indexing, and so forth. In addition to diagrams and example text transformations, the authors use "pseudo-code" to present algorithms in a language-independent manner wherever possible. They also supplement the reading with mg--their own implementation of the techniques. The mg C language source code is freely available on the Web. Alone, this book is an impressive collection of information. Nevertheless, the authors list numerous titles for further reading in selected topics. Whether you're in the midst of application development and need solutions fast or are merely curious about how top-notch information management is done, this hardcover is an excellent investment. --Stephen W. Plain Topics

covered: Text compression models, including Huffman, LZW, and their variants; trends in information management; index creation and compression; image compression; performance issues; and overall system implementation. "This book is the Bible for anyone who needs to manage large data collections. It's required reading for our search gurus at Infoseek. The authors have done an outstanding job of incorporating and describing the most significant new research in information retrieval over the past five years into this second edition."-Steve Kirsch, Cofounder, Infoseek Corporation"The new edition of Witten, Moffat, and Bell not only has newer and better text search algorithms but much material on image analysis and joint image/text processing. If you care about search engines, you need this book: it is the only one with full details of how they work. The book is both detailed and enjoyable; the authors have combined elegant writing with top-grade programming."-Michael Lesk, National Science Foundation"The coverage of compression, file organizations, and indexing techniques for full text and document management systems is unsurpassed. Students, researchers, and practitioners will all benefit from reading this book."-Bruce Croft, Director, Center for Intelligent Information Retrieval at the University of MassachusettsFrom the Back Cover"This book is the Bible for anyone who needs to manage large data collections. It's required reading for our search gurus at Infoseek. The authors have done an outstanding job of incorporating and describing the most significant new research in information retrieval over the past five years into this second edition." Steve Kirsch, Cofounder, Infoseek Corporation"The new edition of Witten, Moffat, and Bell not only has newer and better text search algorithms but much material on image analysis and joint image/text processing. If you care about search engines, you need this book: it is the only one with full details of how they work. The book is both detailed and enjoyable; the authors have combined elegant writing with top-grade programming." Michael Lesk, National Science Foundation"The coverage of compression, file organizations, and indexing techniques for full text and document management systems is unsurpassed. Students, researchers, and practitioners will all benefit from reading this book." Bruce Croft, Director, Center for Intelligent Information Retrieval at the University of MassachusettsIn this fully updated second edition of the highly acclaimed Managing Gigabytes, authors Witten, Moffat, and Bell continue to provide unparalleled coverage of state-of-the-art techniques for compressing and indexing data. Whatever your field, if you work with large quantities of information, this book is essential reading--an authoritative theoretical resource and a practical guide to meeting the toughest storage and access challenges. It covers the latest developments in compression and indexing and their application on the Web and in digital libraries. It also details dozens of powerful techniques supported by mg, the authors' own system for compressing, storing, and retrieving text, images, and textual images. mg's source code is freely available on the Web.