Managing Gigabytes:
Compressing and Indexing documents and images

By Ian H. Witten, Alistair Moffat, and Timothy C. Bell
Second Edition
MORGAN KAUFFMAN, Copyright 1999, 519 pp.
ISBN 1558605703 , List Price $62.95

Review by:
S. V. Nagaraj
svn1999@eth.net

This book published in 1999 is a revised second edition of the version that first appeared in 1994. The authors have incorporated many changes in the second edition by updating various chapters to include the latest developments. The book is concerned with the task of managing large volumes of text and image data amounting to several gigabytes. The fundamental problems addressed in the book include compressing such data and indexing it, in order to enable easy search. Many books look at compressing and indexing as if they are unrelated techniques; however, this book brings out the advantages of combining them in a beneficial way.

The book is meant for a wide variety of readers including software professionals, librarians, and distributors of items such as CD-ROMs. The book has been authored by academics and is therefore well suited for academic use. It may be used for teaching courses in the area of data compression and information retrieval at various levels. An instructor's supplement is also available and includes test questions and review material for use during teaching.

As a supplement to the book, a system known as mg (short for Managing Gigabytes) developed by the authors is available for download freely from the Web site of the authors along with the complete source code in C. The mg system was developed for use on the Unix platform.

The book has two appendices, one of which is a guide to the mg system, whereas, the other is a guide to the New Zealand Digital Library (NZDL, a repository of information freely available on the Web). NZDL makes use of mg as its kernel. The guide to the mg system describes the installation process, a sample storage and retrieval session, utilities for database creation, the process of querying an indexed document collection, managing images, and a suite of three image compression programs.

The book comprises ten chapters. Chapter 1 presents an overview of the book and introduces topics such as document databases, compression, indexes, images of documents and the mg system.

Chapter 2 is about text compression. It discusses topics such as models of the data that needs to be compressed and includes adaptive models. Huffman codes are discussed along with algorithms and data structures for dealing with them. Arithmetic coding is discussed along with techniques for implementing it. Symbol-wise models are introduced and four data compression techniques based on them are discussed. They are Prediction by Partial Matching (PPM), block-sorting compression, Dynamic Markov Compression (DMC) and word-based compression. Dictionary-based compression models such as LZ77, LZ78, and the LZW variant of LZ78 are discussed. Synchronization methods for achieving random access in compressed files are also described. The performance of various

compression techniques is evaluated by using different types of input data.

Chapter 3 is on indexing; the process of creating indexes to aid the process of searching text efficiently by means of keywords. Sample document collections are used to illustrate the indexing complexities associated with different types of data. The application of inverted files for enabling indexing is described. The processes of compressing and indexing inverted files are also discussed. The performance of various index compression methods is evaluated. Signature files and bitmaps are described as two other approaches to indexing. The three indexing methods are then contrasted.

Chapter 4 deals with the subject of querying indexes. Various data structures to store lexicons are discussed. Disk-based lexicon storage is described besides the application of minimal perfect hashing for accessing the lexicon. Methods for querying when query terms are only partially specified are discussed. Techniques for processing Boolean queries are described. Alternatives to Boolean queries are also explored. Interactive and distributed retrieval techniques are discussed and the retrieval efficiency of various methods is assessed.

Chapter 5 is concerned with the arduous task of index construction. Memory-based and sort-based inversion approaches for index construction are described. The application of compression for index construction is discussed. This leads to a technique known as compressed in-memory inversion. The inversion methods are then compared. Techniques for constructing signature files and bitmaps are discussed in addition to methods for dealing with dynamic collections.

Chapter 6 is concerned with image compression. The CCITT fax standard and the JBIG standard for bilevel images are discussed. The GIF and PNG image formats are described. The FELICS and CALIC lossless image compression techniques are discussed along with JPEG-LS (lossless JPEG) technique. JPEG is a lossy compression standard for continuos-tone images.

Chapter 7 is concerned with textual images (images of texts). Both lossy and lossless compression techniques are discussed and their performance is evaluated. The JBIG2 standard for textual image compression is also described.

Chapter 8 deals with techniques for dealing with instances when both text and images occur together in documents. In such cases, it is worthwhile to differentiate between text and images.

Chapter 9 is concerned with the implementation-related aspects of various techniques described in the book. For text compression, the choice of compression model has a considerable impact. Text compression performance is explored focussing on effectiveness, decompression speed and the impact of memory. The implementation of index construction, index compression and query processing are also discussed.

Chapter 10 speculates on the future of techniques for managing gigabytes. The huge impact of the Internet and utilities such as digital libraries on every day life and the need for better data compression techniques is emphasized.

The book includes adequate references to the literature that were current at the time of publication. The handy index aids easy referencing.

Overall, the book is a treasure trove of techniques for compressing and indexing. It will be handy for those interested in indexing and retrieving information from large text databases running to several gigabytes. The style of presentation is laudable and the coverage of topics is impressive. The authors could have included audio and video compression techniques. This good book will be very useful for the intended audience.