Managing Gigabytes: Compressing and Indexing Documents and Images

I.H. Witten, A. Moffat, T.C. Bell

Van Nostrand Reinhold, New York, 1994
ISBN 0-442-01863-0

Errata as at February 26, 1996

Front Matter

page vii, last line: "Index 425" → "Index 423"

Chapter One

Chapter Two

page 25, para 2, line 5: "is is" → "it is"

page 34, Table 2.2: The two 18 bit codes are what `mg` actually produces, but strictly speaking they should be 17 bits long, since 100, 101, 102 etc all appear exactly once in the text. `Mg` extends two codes to reserve one codeword for other purposes, but this is not explained at all, and the example should have been hand edited to avoid confusion. If anything, the example is a little bit too realistic in the context of Chapter Two.

page 42, para 2: The description of PPM should mention the idea of "exclusions". In the example, the context *lie* has occurred 201 times. However, 22 of these occurrences were followed by the letter "s", yet an "s" will never be coded in that context, so it can be *excluded* from the count. This means that only 179 of the occurrences of the context will be used as a sample, and an "r" will have an estimated probability of 19/180. Performing exclusion takes a little extra time, but gives a reasonable payback in terms of compression. There are several other variations of performing sampling that also offer different tradeoffs between speed and compression.

page 43, para 3, line 4: "is the number them that" → "is the number of them that"

page 47, line –8: "such a capital letter following a period" → "such as a capital letter following a period".

page 53, para 3, line 1: "Gzip" → "GZip"

page 64, Table 2.5, line "progp": "43,379" → "49,379"

page 68, Table 2.6: "Mbyte/sec" → "Mbyte/min" twice in the body of the table, and in the caption "Mbyte/second" → "Mbyte/minute" page 70, para 4, line 5: “Santos” $\rightarrow$ “Santis”

page 71, line 11: “Fiala and Greene (1989)” $\rightarrow$ “Fiala and Green (1989)”

Chapter Three

page 89, para starting “Using this method”, line 2: “*hapax legomena*” $\rightarrow$ “*hapax legomenon*”

page 96, line 5: “a such a” $\rightarrow$ “such a”

page 98, line 6: “shows that in fact none is an answer to this query” $\rightarrow$ “shows that only document 2 is an answer to this query”

page 106, para 3, line 9: “the bitstring in Figure 3.7b” $\rightarrow$ “the bitstring in Figure 3.7c”

page 107, Figure 3.7: The coding shown in part (c) cannot be decoded ambiguously. For example, the sequence “1010 0000 0001 0000” would be represented as “1010, 00, 10, 11”, but so would the sequence “1000 0000 0011 0000”. One way to avoid this problem is to add an extra bit to the position codes to say whether or not they are the last one for their subsequence; another is to prefix, at each expansion, the number of non-zero subtrees.

page 108, Figure 3.8: The stemmer that was used for this example is proprietary, and so cannot be distributed with `mg`. Hence, if you try this experiment using the distributed `mg` software (which contains a much simpler stemmer) you are likely to get different stemmed forms.

page 115, para starting “The tradeoff”: “MacKenzie” $\rightarrow$ “McKenzie”

page 115, para starting “Lovin”: “Lovin” $\rightarrow$ “Lovins”.

Chapter Four

page 160, line “and the code c”: “code c for integer x” $\rightarrow$ “code c for x”

page 173, line 3: “Baye’s Theorem” $\rightarrow$ “Bayes’ Theorem”

page 174, line -5: “Sparc-Jones” $\rightarrow$ “Sparck-Jones”

Chapter Five

pages 189–195: We have improved the algorithm described in this subsection, and the description here, while correct, is no longer up to date. Contact Alistair Moffat for more details if you are interested.

page 195, para 1: “400×50 Kbytes $\times 1/2 = 1$ Mbyte” $\rightarrow$ “400×50 Kbyte $\times 1/2 = 10$ Mbytes”

page 201, formula 2, line 2: “$10f$” $\rightarrow$ “$10ft_r$”

page 204, Table 5.5: On the line "Sort-based, multi-way, in-place", "141" → "150"

Chapter Six

page 245, line 2: "Recall from Section 3.3 that the Golomb code has a" → "In Section 3.3 we described the Golomb code. It is controlled by a". (Section 3.3 is marked by a gray bar, and it seems unfair to ask the reader to recall something they may not have read.)

Chapter Seven

page 294, line 1: "Mohuiddin" → "Mohiuddin"

Chapter Eight

Chapter Nine

page 349, para 2, line 8: "52" → "53", twice. The Sun uses a floating point format with a 52-bit mantissa, but a normalised value always has a leading "1" bit and this is not stored, hence, to be completely accurate, there are 53 bits of integer accuracy available. The same correction applies also to page 351, para 3, line 5.

page 351, para 1: Length-limited prefix codes were not discussed in detail, as the existing algorithms were impractical for use on large alphabets because of their requirement for large amounts of memory. However, since the book went to press we have devised improved methods for implementing length-limited codes, and it is now possible to calculate them in pretty much the same space and time as ordinary Huffman codes. Hence, much of the material in this section starting on page 345 should now be replaced by a description of how to implement length-limited Huffman codes. For details of the new algorithms, contact Alistair Moffat. We have also developed new space-efficient techniques for calculating Huffman codes.

page 355, para 2, line 6: "*huffword* decodes faster than the *pack*" → "*huffword* decodes faster than *pack*"

page 367, section 9.5, line 10: "Table 4.8 on page 149" → "Table 4.9 on page 156"

page 372, line 3: "used is as an aid" → "used as an aid"

Chapter Ten

Appendix

page 391, last line: `Mg` is also available as a `gzip` archive. In this case, copy the file `mg.tar.gz` and use the command

```
gunzip < mg.tar.gz | tar xf -
```

pages 393–398: Since completing the book we have developed an X-windows interface. This is not described at all in the appendix. To try this interface, execute `xmg` and then click on the help button. Queries are typed in the top text-box, and executed when enter is typed. Inspect documents by clicking on their first line.

There is a manual page for `xmg` (as there is for all of the `mg` commands), type `man xmg` as a Unix shell command.

page 394, Figure A.1: The output format of `mgstat` has changed since this figure was prepared. Also, note that the different stemmer included with `mg` will give rise to different statistics for the `alice` collection. You are unlikely to be able to reproduce the numbers in Figure A.1(a).

page 406, Table A.4, line 1: A further `.set mode` option has been added:

```
.set heads_length 20
.set mode heads
```

displays just the first 20 characters of each answer to subsequent queries, with whitespace characters compacted. To select and examine one returned document in full, use the document number (33, say) from the desired line and return to `text` mode and change to `docnums` as the query mode:

```
.set query docnums
.set mode text
33
```

This mode is used to good effect in `xmg`.

Bibliography

page 413, line 1: "A. De Santos" shouldbe "A. De Santis"

page 417, "Morgenstern, H.B. and S. Morgenstern" "Morgenstern, H.B." → "Barlow, H."

page 418, ref "Robertson" "Sparck Jones" → "Sparck-Jones"

Index