Implementing a Novel Data Structure for Maintaining Cumulative Frequency of Symbols

Abstract

A new data structure, namely "cumulative frequency matrix (CFM)", is proposed here for maintaining cumulative frequencies. For an order-0 model having 256 symbols, CFM is a 2-D array of 16 rows and 16 columns. Two nibbles, say L for left and R for right, of a byte symbol represents row and column dimensions respectively. Matrix element (L, R) represents cumulative frequency of symbol with right nibble as R among symbols with left nibble as L. Within row, it stores cumulative frequency of symbols with right nibble varying from 0 to 15. Adaptive arithmetic coding is a lossless data compression method. It needs to update cumulative frequencies at runtime. Various algorithms for maintaining cumulative frequencies, computing cumulative frequency interval etc. are discussed here. Practical implementation shows that proposed data structure is simpler as well as efficient as compared to other data structures in use.

References

- Peter M. Fenwick, "A New Data Structure for Cumulative Frequency Tables; Software – Practice and Experience, Vol 24(3), 327-336 (March 1994)
- Ian H. Witten, Alistair Moffat, Timothy C. Bell, "Managing Gigabytes-Compressing and Indexing Documents and Images; 2nd edition, Morgan Kaufmann Publishers
- Ida Mengyi Pu, Fundamental Data Compression, Butterworth-Heinemann, 2006
- Algorithm Tutorials available at http://community. topcoder. com

**Index Terms**

Computer Science

Algorithms
Keywords
Adaptive Arithmetic Coding  Data Compression  Cumulative Frequency Maintenance  Data Structure  Algorithm