Abstract

Byte pair encoding (BPE) algorithm was suggested by P. Gage to achieve data compression. It encodes all instances of most frequent byte-pair using zero-frequency byte in the source data. This process is repeated for maximum m possible number of passes until no further compression is possible, either because there are no more frequently occurring byte pairs or there are no more unused zero-frequency bytes to represent pairs. It writes out substitution information before the encoded data in each pass. This algorithm is very time consuming as it requires to determine most frequent byte-pair in each pass before starting substitution. We have proposed k-pass byte-pair transformation algorithm where k may be very very small as compared to maximum possible passes m. Our aim is to minimize the compression time and achieve equivalent compression rate. Proposed algorithm transforms half of the possible most-frequent byte pairs in each pass except the last. In the last pass, it transforms all remaining possible byte-pairs. This reduced number of passes save the time taken in computing frequency of byte-pairs in maximum m passes. Experimental results have shown that proposed algorithm had taken 3. 213, 9. 794, 13. 324, 16. 323, 22. 388 seconds with 1, 2, 3, 4 and 6 passes respectively as compared to 295. 642 seconds of m-passes. Compression rate achieved due to transformation is 14. 72%, 20. 12%, 21. 89%, 22. 67% and 22. 96% with 1, 2, 3, 4 and 6 passes respectively as compared to 25. 55% using maximum m-passes. As the
number of passes increases, compression is better with increased execution time. Our aim of achieving speed is achieved with little loss in compression rate.

References

- Altan Mesut, Aydin Carus, "ISSDC: Digram Coding Based Lossless Data Compression Algorithm"; Computing and Informatics, Vol. 29, pp. 741–754, 2010
- Sayood Khalid, "Introduction to Data Compression"; 2nd edition, Morgan Kaufmann, 2000
- mattmahoney.net/dc/bpe2v2.cpp
- Jyotika Doshi, Savita Gandhi, "Quad-Byte Transformation as a Pre-processing to Arithmetic Coding"; International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 12, December 2013, e-ISSN: 2278-0181

Index Terms

Computer Science   Signal Processing

Keywords

Data Compression   better compression rate   byte-pair data transformation   substitution using zero-frequency byte symbols