

AN IMPROVED METHOD FOR LOSSLESS DATA COMPRESSION

Yuhua Bai and Todor Cooklev

School of Engineering, San Francisco State University, San Francisco CA 94132, USA

This paper describes an improved lossless data compression scheme. The proposed scheme contains three innovations: first, an efficient algorithm is introduced to decide when and how to switch from transparent mode to compressed mode, second, a temporary buffer is introduced at the encoder, and third, an approach to decide when to discard the entire dictionary is advanced. According to the developed method the changes are only at the transmitter and any V.42bis-compatible receiver can be used as a decoder. Therefore devices using V.42bis can use the proposed method after a firmware upgrade.

The algorithm specified in V.42bis operates in one of two modes: transparent (T) mode and compressed (C) mode. In transparent mode, symbols are sent uncompressed, while in compressed mode symbols are compressed using the LZW algorithm and represented by codewords. Note that the compression ratio achieved by V.42bis operating in compressed mode depends on the compressibility of the data. Therefore, it is advantageous to switch from compressed mode to transparent mode when incompressible data is encountered. However V.42bis does not specify when and how the encoder can decide to switch from compressed mode to transparent mode and vice-versa. The standard also does not describe when and how an encoder can decide to drop the entire dictionary. We analyze the cost to switch modes. The cost to switch from T mode to C mode is $C_{TC} = 16$ bits.

The cost to switch from C to T mode is $C_{CT} = CW + CW + Z$, where CW is the codeword size for current processed character in bits, and the second CW is the codeword size for the enter transparent mode (ETM) control word. Z is the number of zeros that must be transmitted to recover the byte alignment and has an expected value of $E(Z) = 7/2$. To make a decision when to switch modes, we introduce two look-ahead buffers, B_C and B_T , for each mode of operation of the encoder. Regardless of which mode the encoder is in, the output of both modes of operation is written to the corresponding look-ahead buffer. The threshold to switch from T mode to C mode is determined to be $NB_C / NB_T \leq 1 - (C_{TC} + p_{ct} C_{CT}) / 8n$, where NB_C and NB_T are the number of bits in the two buffers. The threshold to switch from C mode to T mode is determined to be $NB_C / NB_T \geq 1 + (C_{CT} + p_{tc} C_{TC} - p_{ct} C_{CT}) / (8n)$. The simulation results demonstrate that the proposed method achieves higher compression ratios in most cases. Another goal of the work presented in this paper is to analyze what is the improvement obtained after the dictionary is reset and when is a good time to discard the dictionary. Our results for different file types show that consistently the compression ratio before dictionary reset is 0.87 and after dictionary reset is 1.07, an increase of 22.6 %. It is noted that V.44 is a newer compression standard, based on a different compression algorithm. While our results do not apply directly to V.44, they may be used after appropriate modifications.