Low Complexity Parametrized Codes for LZ77 Compression

Péter A. Felvégi

The family of LZ77 derived compression algorithms is rather long. Newer and newer variants have been proposed since the publication of the original algorithm in 1977. Of these, some have focused on how to improve the match finding part in terms of temporal and spatial complexity, the others delt with the coding part of the algorithm. Nowadays, the most widely used popular compression programs also use some variation of the LZ77 algorithm (such as zip, gzip, arj, rar, zlib). These programs can achieve good compression ratios in acceptable time, though they sacrifice some compression for being ’fast’. The resource constraints are usually the same or similar for compression and decompression, these programs are usually run on PC’s or workstations.

In some applications the resource constraints for the compression and decompression parts might be remarkably different. This is the case generally when compression takes place only once and decompression many times, e.g.: at embedded systems, static databases, executable files, distribution CD’s, etc. For such applications, the main aim is to maximize compression while keeping the decompressor complexity as low as possible in terms of time and maybe for memory requirements, too. For the compression part, the constraints are far less strict. Since compression takes place only once, it can be done off-line using reasonable amount of computing power and memory.

With the above assumptions, the LZ77 algorithm can be improved in two ways: better parsing during the string matching part; and better coding of the literals, distance and length values. In this article we will focus on efficiently encoding the match lengths.

The range of the match lengths is between the minimal allowed (typically 2,3) and the maximal allowed (typically 256-64K) values. On this range, the distribution of the values tend to be asymptotically decaying, thus entropy coding offers a gain against the flat binary code. We omit adaptive codes and also omit static Huffman coding, because we want to keep decompressor complexity at a minimum.

Simple codes are already proposed by several people (Elias, Golomb, Rice, Fiala, et al.) for representing integer numbers of an assumed distribution. In this article I will present a family of codes that are parametrized by a few integers, and are capable to adapt better to the actual distribution of the values, thus increasing the coding efficiency. With the proper set of parameters, some of the cited codes can be achieved, too. The drawback of the parametrized codes is that the parameters must be optimized first, thus compressor complexity is usually higher than with the other codes – which was the original assumption.

I will present the comparison of these codes to the others, and also to Huffman codes and to the theoretical optimum based on the entropy. The work is still in progress, other parametrized codes may be found later for a special purpose/distribution.