Sidharth Ghoshal

A method of compression through variability exploitation and bit substitution

Consider a file represented as a stream of bits

$$File \in \{0,1\}^n$$

The following compression algorithm can be implemented

Select an integer k

Create a 2 hashtables H_1, H_2 with each of their bucket values initialized to -1 and an index value $e = 0$

Parse the file k-bits at a time, for each cluster of k bits Q,

Use it as a key in H_2 and see if a non-negative integer is returned. If found the cluster has already been encountered

If -1 is returned store Q in Hash table H_1 with key e, stores e in Hash table H_2 with key Q

increment e

Once complete, take the contents of Hash table H1 and strip all the buckets with -1 (transform it into an array). Denote the length of the table L

Treat the file as a number in base L (whereas the individual bit sequences of k bits at a time relate to different integers via the table H1)

Convert this number to base 2^k with position holders represented by all the possible bit strings of size k (ordered as numbers in binary)

The newly converted number will have size less than or equal to the original file. Take the compressed file (which is the converted number) and the associate table, and store.

To decompress, it can be converted from base 2^k to base L and then substitutions can be made to recover the original file.

The order of compress can be calculated as follows:

$$k = \text{size of parse}$$

$$unique\ sequences\ in\ original\ file = L$$

$$n = original\ file\ size$$

$$Compressed\ File\ Size\ =\ \log_{2^k}(L^n) +\ kL$$

The first term is the size of the compressed file, the second term is the size of associated hash table for conversion back. The challenge is in picking an appropriate value k which is both small but also selected such that not all permutations of bit sequences of that size have appeared. We note that if

For a given size k the size of all possible bit sequences of that size is

$$2^k k$$

Thus if

$$2^k k > n$$

We are guaranteed that some compression will occur.