Start/Stop Codes

Steven Pigeon
Université de Montréal
pigeon@iro.umontreal.ca

Abstract

Fiala and Greene's (*start*, *step*, *stop*) codes offer a convenient and easy way to implement variable length codes, but hardly take distribution information into account [Fiala89]. We present a similar method that uses more information on the distribution to compute an efficient code, at a very low cost in memory for code description.

Start/Stop codes resemble Golomb codes [Golomb66] in that they are composed of a prefix that encodes the length of the integer and a suffix that encodes the integer itself. Golomb codes ask for the computation of a fixed constant, b, the order of the code, so that each prefix bit has a maximal information contents, under a geometrical distribution hypothesis. The prefix is the unary coding of the length of the integer expressed as a multiple of b. Start/Stop codes, given eventually non-increasing distributions on a range, allow each of the prefix bits to be optimized separately, resulting in a set of segment lengths $\{m_0, m_1, \ldots, m_k\}$ rather than an only parameter b. The segment lengths for a random source X are chosen such that $\mathrm{P}(X < 2^{m_0}) \approx \frac{1}{2}$, $\mathrm{P}(X < 2^{m_0} + 2^{m_0+m_1} \mid X \geq 2^{m_0}) \approx \frac{1}{2}$, $\ldots$, $\mathrm{P}(X < \sum_{i=0}^{k} 2^{\sum_{j=0}^{i} m_j} \mid X \geq \sum_{i=0}^{k-1} 2^{\sum_{j=0}^{i} m_i}) \approx \frac{1}{2}$, $\ldots$, and in a way that minimizes first order entropy and average Start/Stop code length discrepancy.

Start/Stop codes can be modified to be used as an universal code for the integers, even taking into account a distribution. A threshold i and a set of segment lengths $\{m_0, m_1, \ldots, m_{i-1}, m_i, m_i, \ldots\}$ can be used to code arbitrarily large integers. The first i segment lengths are optimized, and the remaining are optimized as a Golomb code, letting $m_i = b$, as if the tail of the distribution were geometric. The threshold is chosen as to have acceptable compression loss.

Start/Stop code can be encoded with all prefix bits preceeding the bits of the integer to code, giving codes of the form $111\ldots10\langle m_0\rangle\langle m_1\rangle\ldots\langle m_s\rangle$, where $\langle x\rangle$ represent a string of x bits. Or they can be in a way that prefix bits are paired with their associated segments, giving codes of the form $1\langle m_0\rangle 1\langle m_1\rangle\ldots 0\langle m_s\rangle$. If the m_k are chosen to be of the form $2^l - 1$, the prefix bit and segment bits pairs are aligned on machine words boundaries and efficient coding and decoding algorithms and be applied, albeit to the expense of compression efficiency.

Analytic and empirical evidence show that the Start/Stop codes have good efficiency against distribution gathered from files and standard distributions.

[Elias75] P. Elias, *Universal Codeword Sets and Representations of the Integers*, IEEE Trans. on Inform. Theory, IT-21 #2, pp. 194-203, March 1975

[Fiala89] E.R. Fiala, D.H. Greene, *Data compression with finite windows*, CACM, v32 #4, pp. 490-505, April 1989

[Golomb66] S.W. Golomb, *Run Length Encodings*, IEEE Trans. on Inform. Theory, IT-12 #3, pp. 399-401, March 1966