

The Discreet Charm of Multi-Pattern Codes

(*Abstract*)

Igor Zavadskyi

Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
2d Glushkova ave.
`ihorzavadskyi@knu.ua`

Probably the most important trade-off in data compression is between the compression ratio and code processing speed. When it comes to compression ratio, methods approaching the theoretical bound of entropy encoding have been known for decades, such as optimal arithmetic encoding and quasi-optimal Huffman codes. The latter codes can be processed times faster, while their compression efficiency can vary from optimal to significantly suboptimal depending on the properties of the source alphabet. In 2014, J. Duda et al. proposed the encoding based on Asymmetric Numeral Systems, providing a compromise solution nearly as good as arithmetic encoding in terms of compression ratio and nearly as fast as Huffman codes.

Yet, these impressive solutions tend to overshadow codes that prioritize fast decoding, leaving many intriguing questions unanswered in this less-explored domain. Can we develop a code that can be processed significantly faster than Huffman codes? If so, what are the trade-offs? What specific structural properties of a codeword set influence the speed of decoding? For instance, a step structure of codeword length distribution may accelerate code processing. In byte-aligned codes (ETDC, SCDC, or RPBC), invented in the early 2000s, codewords are composed of whole bytes and thus can be processed easily and quickly. However, this is achieved at the cost of a 12–15% loss in compression ratio. Fibonacci codes are not so stepped, thus becoming denser but slower.

Our presentation delves into a series of recently developed multi-pattern variable-length data compression codes. While trading a small percentage of compression ratio, these codes offer an order-of-magnitude acceleration of code processing, surpassing both byte-aligned and Fibonacci codes in terms of compression efficiency and processing speed. We explore the structural, algorithmic, and technical aspects of fast decoding and showcase several other benefits of these innovative codes, including the potential for fast Boyer-Moore-style search in a compressed file and the ability to represent integer sequences in a space-efficient manner with nearly constant time direct access.