Recursive Order-Based Compression for Permutable Lists

Author: Bernardo Machado Ferreira
Instituto Superior de Engenharia do Porto (ISEP), Politécnico do Porto
Supervisor: Nuno Escudeiro

Abstract: Permutable lists, sets, and multisets can be looked at in two ways: as unordered collections, whose entropy is usually captured by treating them as opaque binary strings once their order is set aside, or as sequences, whose order is encoded separately from the elements it arranges. Between these two views lies an open question — whether the order in which elements appear can itself be read as information, and folded back into the representation of the data rather than discarded or kept apart from it. This dissertation takes up that question, proposing a recursive order-based compression method that bridges sequence representations and multiset entropy theory, in which a multiset is represented as a list of its original, or compressed, elements. At the heart of the method is a new permutation encoder that lets the order of a list's elements carry part of the list's own information; this order-derived information is then recursively unfolded into further elements, each sublist read against the order of the one that came before it. This recursive folding is shown to yield savings that are asymptotically optimal, converging to the entropy of the underlying permutation as the list grows, all while the resulting representation remains directly indexable, with random access preserved throughout. The method is realized through a Fenwick tree for O(log K) rank queries, together with worst-case and average-case partitioning strategies, brought together in an end-to-end encoding and decoding pipeline, and its practical compression behavior is characterized experimentally. The resulting contribution is a general encoding scheme that draws order-based and multiset compression into a single, coherent framework — a representation that is both compact and immediately usable, without giving up one quality for the other.

Keywords: compression, order, permutations, lists, sets, multisets, random access