Title: Suffix Graphs and Lossless Data Compression

Author: Martin Senft

Department: Department of Software and Computer Science Education

Supervisor of the doctoral thesis: doc. RNDr. Tomáš Dvořák, CSc., Department of Software and Computer Science Education

Abstract: Suffix tree and its variants are widely studied data structures that enable an efficient solution to a number of string problems, but also serve for implementation of data compression algorithms. This work explores the opposite approach: design of compression methods, based entirely on properties of suffix graphs. We describe a unified construction algorithm for suffix trie, suffix tree, DAWG and CDAWG, accompanied by analysis of implicit suffix link simulation that yields two practical alternatives. Since the compression applications require maintaining text in the sliding window, an in-depth discussion of sliding suffix graphs is needed. Filling gaps in previously published proofs, we verify that suffix tree is capable of perfect sliding in amortised constant time. On the other hand, we show that this is not the case with CDAWG, thus resolving a problem of Inenaga et al. Building on these investigations, we describe a family of data compression methods, based on a description of suffix tree construction for the string to be compressed. While some of these methods resemble classical lossless compression like PPM or dictionary techniques, others appear to be brand-new. Our algorithms are illustrated with a proof-of-concept implementation and experimental evaluation on standard data corpora.

Keywords: suffix tree, CDAWG, construction, sliding window, data compression