September 2010

MADALGO seminar by Philip Bille, Technical University of Denmark

Random Access to Grammar Compressed Strings

Abstract:

Grammar based compression, where one replaces a long string by a small context-free grammar that generates the string, is a simple and powerful paradigm that captures many of the popular compression schemes, including the Lempel-Ziv family, Run-Length Encoding, Byte-Pair Encoding, Sequitur, and Re-Pair. In this paper, we present a novel grammar representation that allows efficient random access to any character or substring without decompressing the string.

Let S be a string of length N compressed into a context-free grammar $\mathcal{S}$ of size n. We present two representations of S achieving $O(\log N)$ random access time, and either $O(n \cdot \alpha_k(n))$ construction time and space on the pointer machine model, or $O(n)$ construction time and space on the RAM. Here, $\alpha_k(n)$ is the inverse of the k^{th} row of Ackermann's function. Our representations also efficiently support decompression of any substring in S: we can decompress any substring of length m in the same complexity as a single random access query and additional $O(m)$ time. Combining these results with fast algorithms for uncompressed approximate string matching leads to several efficient algorithms for approximate string matching on grammar compressed strings without decompression. For instance, we can find all approximate occurrences of a pattern P with at most k errors in time $O(n\,(min\,|P|k, k^4 + |P| + \log N) + occ)$, where occ is the number of occurrences of P in S. Finally, we are able to generalize our results to navigation and other operations on grammar-compressed *trees*

All of the above bounds significantly improve the currently best known results. To achieve these bounds, we introduce several new techniques and data structures of independent interest, including a predecessor data structure, two "biased" weighted ancestor data structures, and a compact representation of heavy-paths in grammars.

Joint work with Gad M. Landau, Rajeev Raman, Kunihiko Sadakane, Srinivasa Rao Satti, and Oren Weimann