Problem Description: The string indexing with compressed pattern problem

Mathias Søndergaard, s174426

September 5, 2022

Problem definition

Given two strings, S and P, the string indexing problem consists of finding all starting positions of P in S. Commonly P is known as the pattern with $|S| > |P|$. In this thesis we shall investigate the problem where P is compressed, which is known as the string indexing with compressed pattern problem (SICP). We will use the LZ77 compression scheme to compress P, with $LZ(P)$ denoting the phrases obtained by compressing P. A typical use case which is affected by the SICP problem is in the client-server scenario. Here the server contains S. In order to save bandwidth, the client sends a request which includes $LZ(P)$, thus having the server do the indexing.

Previous work

A naive solution would be to decompress $LZ(P)$, transforming the SICP problem instance to a string indexing instance. Let $m = |P|$ and let occ denote the number of occurrences of P in S. The naive method results in $O(m + occ)$ time, as $LZ(P)$ has to be decompressed and the indexing is done using a suffix tree. Both require $O(m)$ and reporting occ elements require $O(occ)$. $O(n)$ space is required for the suffix tree, with $n = |S|$. Novel solutions exist which does not decompress $LZ(P)$ while obtaining better time bounds. These are covered in [2], with the main finding being a data structure using $O(n)$ space and $O(z + log(m) + occ)$ time. Here $z = |LZ(P)|$.

Goals of the project

The primary goal of this project is to review and implement algorithms / data structures for the string indexing with compressed pattern problem. These will be based on state of the art theory, given in [2]. Moreover, they will be tested and bench-marked, investigating whether the theory holds in practice. To reach this goal, other data structures and algorithms will be covered and implemented, including Nearest common ancestor, Longest common prefix, LZ77, Suffix trees, ART-decomposition ([1]) and their optimal prepossessing algorithms.

References

[1] Stephen Alstrup, Thore Husfeldt, and Theis Rauhe. “Marked ancestor problem”. In: (1998).

[2] Philip Bille, Inge Li Gørtz, and Teresa Anna Steiner. “String Indexing with Compressed Patterns”. In: (2020).