Enhanced and Extended Suffix Arrays

Adrian Regenfuß

June 16, 2020

Abstract

In this report, I review the literature on enhanced and extended suffix arrays in the context of searching long strings. I examine the different algorithms used for both constructing enhanced and extended suffix arrays and for using them in searching long strings. In the end, I compare enhanced and extended suffix arrays with suffix arrays and suffix trees.

Introduction

Finding the occurences of one string in another string, longest repeated substrings and longest shared substrings of two different strings are fundamental problems for many kinds of computing systems.

As a result, many different algorithms have been developed for these kinds of problems: For finding the occurences of one string in another one the naive algorithm and the Boyer-Moore algorithm, and for all three of these problems (and more) three different data structures: the suffix tree, the suffix array Manber and Myers 1993 and the enhanced suffix array.

Suffix trees, suffix arrays and enhanced suffix arrays have the disadvantage of requiring to be constructed for a specific string, which has time and space requirements. Because of this, they are better suited for tasks where unchanging strings have to be searched or matched repeatedly, although there has been some work to extend the suffix array to dynamic strings.

Since searching and matching very long unchanging strings is very common in genome analysis, it doesn't surprise that both suffix arrays and enhanced suffix arrays were developed in that context.

Different String Matching Problems

A plethora of different string matching problems have been identified by computer scientists, for many of which suffix arrays and enhanced suffix arrays are useful.

Searching

Suffix Trees

Suffix Arrays

Construction

Searching

Enhanced and Extended Suffix Arrays

Enhanced suffix arrays were first proposed in Abouelhoda et al. 2004 as an improvement over normal suffix arrays. An enhanced suffix array contains a suffix array together with the LCP-array of the string, and sometimes a Burrows-Wheeler transformation and an inverse of the suffix table.

suftab

lcptab

bwttab

suftab^{-1}

LCP-Interval Trees

Searching

Finding Maximal and Supermaximal Repeats

Comparison

Applications

Conclusion

References

Udi Manber and Gene Myers. Suffix arrays: a new method for on-line string searches. *siam Journal on Computing*, 22(5):935–948, 1993.