Matching and Compression of Strings with Automata and Word Packing

Dynamic Relative Compression, Dynamic Partial Sums, and Substring Concatenation

Given a static reference string $R$ and a source string $S$, a relative compression of $S$ with respect to $R$ is an encoding of $S$ as a sequence of references to substrings of $R$. Relative compression schemes are a classic model of compression and have recently proved very successful for compressing highly-repetitive massive data sets such as genomes and web-data. We initiate the study of relative compression in a dynamic setting where the compressed source string $S$ is subject to edit operations. The goal is to maintain the compressed representation compactly, while supporting edits and allowing efficient random access to the (uncompressed) source string. We present new data structures that achieve optimal time for updates and queries while using space linear in the size of the optimal relative compression, for nearly all combinations of parameters. We also present solutions for restricted and extended sets of updates. To achieve these results, we revisit the dynamic partial sums problem and the substring concatenation problem. We present new optimal or near optimal bounds for these problems. Plugging in our new results we also immediately obtain new bounds for the string indexing for patterns with wildcards problem and the dynamic text and static pattern matching problem. Subsequence Automata with Default Transitions

Let $S$ be a string of length $n$ with characters from an alphabet of size $\sigma$. The subsequence automaton of $S$ (often called the directed acyclic subsequence graph) is the minimal deterministic finite automaton accepting all subsequences of $S$. A straightforward construction shows that the size (number of states and transitions) of the subsequence automaton is $O(n)$ and that this bound is asymptotically optimal. In this paper, we consider subsequence automata with default transitions, that is, special transitions to be taken only if none of the regular transitions match the current character, and which do not consume the current character. We show that with default transitions, much smaller subsequence automata are possible, and provide a full trade-off between the size of the automaton and the delay, i.e., the maximum number of consecutive default transitions followed before consuming a character. Specifically, given any integer parameter $k$, $1 < k$, we present a subsequence automaton with default transitions of size $O(nk \log k)$ and delay $O(\log k)$. Hence, with $k = 2$, we obtain an automaton of size $O(n \log \log n)$ and delay $O(\log n)$. At the other extreme, with $k = \sigma$, we obtain an automaton of size $O(n)$ and delay $O(1)$, thus matching the bound for the standard subsequence automaton construction. Finally, we generalize the result to multiple strings. The key component of our result is a novel hierarchical automata construction of independent interest. Deterministic Indexing for Packed Strings

Given a string $S$ of length $n$, the classic string indexing problem is to preprocess $S$ into a compact data structure that supports efficient subsequent pattern queries. In the deterministic variant the goal is to solve the string indexing problem without any randomization (at preprocessing time or query time). In the packed variant the strings are stored with several character in a single word, giving us the opportunity to read multiple characters simultaneously. Our main result is a new string index in the deterministic and packed setting. Given a packed string $S$ of length $n$ over an alphabet $\Sigma$, we show how to preprocess $S$ in $O(n)$ (deterministic) time and space $O(n)$ such that given a packed pattern string of length $m$ we can support queries in (deterministic) time $O(m + \log m + \log \log m)$, where $m = \omega = \log$ is the number of characters packed in a word of size $\omega = \log n$. Our query time is always at least as good as the previous best known bounds and whenever several characters are packed in a word, i.e., $\log \omega$, the query times are faster. Dynamic Partial Sums in Constant Time and Succinct Space with the Ultra Wide Word-RAM Model

The dynamic partial sums problem is to dynamically maintain an array of $n$ integers while supporting efficient access, update and partial sums queries. This classic problem, and its variations, are very well studied in many different computational models [Fre82,FS89,Fen94, HSS11, HR03, HRS96, RRR01, PD04]. We solve the partial sums problem in the ultra wide word-RAM model, recently introduced by Farzan et al. [FLONS15], where we, in constant time, are allowed to manipulate words of size $w2$ and access $w$ memory locations. Farzan et al. [FLONS15] additionally gave a solution to the dynamic partial sums problem by simulating the RAMBO model to obtain a result by Brodnik et al. [BKMN06]. In this paper we present an improved solution to the dynamic partial sums problem in the ultra wide word-RAM model that supports all operations in either constant or $O(\log \log n)$ time, depending on whether we allow multiplication, and succinct space. We pose as an open problem whether it is possible in the ultra wide word-RAM model to additionally support the classic select operation in constant time.

General information

State: Published
Organisations: Department of Applied Mathematics and Computer Science, Algorithms and Logic
Contributors: Skjoldjensen, F. R.
Number of pages: 80
Publication date: 2017

Publication information

Publisher: DTU Compute
Original language: English
(DTU Compute PhD-2017, Vol. 446).
Electronic versions:
phd446_Skjoldjensen_FR.pdf
Research output: Research › Ph.D. thesis – Annual report year: 2017