Algorithms and Data Structures for Strings, Points and Integers - DTU Orbit (10/06/2019)

Algorithms and Data Structures for Strings, Points and Integers: or, Points about Strings and Strings about Points

This dissertation presents our research in the broad area of algorithms and data structures. More specifically, we show solutions for the following problems related to strings, points and integers. Results hold on the Word RAM and we measure space in w-bit words.

Compressed Fingerprints. The Karp–Rabin fingerprint of a string is a useful type of hash value that has multiple applications due to its strong properties. Given a string S of length N compressed into a straight line program (SLP) of size n, we show a O(n) space data structure that supports fingerprint queries, retrieving the fingerprint of any substring of S. Queries are answered in O(lg N) time. If the compression is a Linear SLP (capturing LZ78 compression and variations), we get O(lg lg N) query time.

Our structure matches the best known query time bound for random access in SLPs, and is the first for general (unbalanced) SLPs that answers fingerprint queries without decompressing any text. We also support longest common extension queries, returning the length ℓ that the substrings from two given positions in S are equal. Answers are correct w.h.p. and take time O(lg N lg ℓ) and O(lg lg N + lg ℓ lg lg ℓ) for SLPs and Linear SLPs, respectively.

Dynamic Compression. In the dynamic relative compression scheme, we compress a string S of length N into n substrings of a given reference string of length r. We give data structures that maintain an asymptotically optimal compression in the scheme and support access, replace, insert and delete operations on S. Our solutions support each operation in O(lg n/ lg lg n) time and O(n+ r) space; or O(lg n/ lg lg n) time and O(n+ r lg ℓ) space. They can be naturally generalized to compress multiple strings.

Our solutions obtain almost-optimal bounds, and are the first to dynamically maintain a string under a compression scheme that can achieve better than entropy compression. We also give improved results for the substring concatenation problem, and an extension of our structure can be used as a black box to get an improved solution to the previously studied dynamic text static pattern problem.

Compressed Pattern Matching. In the streaming model, input data flows past a client one item at a time, but is far too large for the client to store. The annotated streaming model extends the model by introducing a powerful but untrusted annotator (representing "the cloud") that can annotate input elements with additional information, sent as one-way communication to the client. We generalize the annotated streaming model to be able to solve problems on strings and present a data structure that allows us to trade off client space and annotation size. This lets us exploit the power of the annotator.

In compressed pattern matching we must report occurrences of a pattern of length ℓ in a text compressed into n phrases (capturing LZ78 compression and variations). In the streaming model, any solution to the problem requires Ω(n) space. We show that the problem can be solved in O(lg n) client space in the annotated streaming model, using O(lg n) time and O(lg n) words of annotation per phrase. Our solution shows that the annotator lets us solve previously impossible problems, and it is the first solution to a classic problem from combinatorial pattern matching in the annotated streaming model.

Pattern Extraction. The problem of extracting important patterns from text has many diverse applications such as data mining, intrusion detection and genomic analysis. Consequently, there are many variations of the pattern extraction problem with different notions of patterns and importance measures. We study a natural variation where patterns must 1) contain at most k different characters, and 2) have at least q occurrences. Both k and q are input parameters.

We show how to extract such patterns and their occurrences from a text of length n in O(nk+q3occ) time and space, where occ is the total number of pattern occurrences. Our bound is the first output-sensitive solution for any approximate variation of the pattern extraction problem, with all previous solutions requiring Ω(n2) time per reported pattern. Our algorithm is relatively simple, but requires a novel analysis technique that amortizes the cost of creating the index over the number of pattern occurrences.

Compressed Point Sets. Orthogonal range searching on a set of points is a classic geometric data structure problem. Given a query range, solutions must either count or report the points inside the range. Variants of this problem has numerous classic solutions, typically storing the points in a tree.

We show that almost any such classic data structure can be compressed without asymptotically increasing the time spent answering queries. This allows us to reduce the required space use if the point set contains geometric repetitions (copies of equal point set that are translated relative to each other). Our result captures most classic data structures, such as Range Trees, KD-trees, R-trees and Quad Trees. We also show a hierarchical clustering algorithm for ensuring that geometric repetitions are compressed.

Points with Colors. Colored orthogonal range searching is a natural generalization of orthogonal range searching which allows us to perform statistic analysis of a point set. We must store n points that each have a color (sometimes called a category) and support queries that either count or report the distinct colors of the points inside a query range.

We show data structures that support both types of queries in sublinear time, storing two-dimensional points in linear space and high-dimensional points in almost-linear space. These are the first (almost) linear space solutions with sublinear query time. We also give the first dynamic solution with sublinear query time for any dimensionality. Previous solutions answer queries faster, but require much more space.

Points with Weights in Practice. If points are each assigned a weight, it is natural to consider the threshold range counting...
problem. A data structure must store the points and be able to count the number of points within a query range with a weight exceeding some threshold. This query appears naturally in a software system built by Milestone Systems, and allows detecting motion in video from surveillance cameras.

We implement a prototype of an index for 3-dimensional points that use little space and answers threshold queries efficiently. In experiments on realistic data sets, our prototype shows a speedup of at least a factor 30 at the expense of 10% additional space use compared to the previous approach. An optimized version of our proposed index is implemented in the latest version of the Milestone Systems software system.

Finger Predecessor. The predecessor problem is to store a set of n integers from a universe of size N to support predecessor queries, returning the largest integer in the set smaller than a given integer q. We study a variation where the query additionally receives a finger to an integer ℓ in the set from which to start the search. We show a linear space data structure that answers such finger predecessor queries in $O(lg \lg |ℓ - q|)$ time. This generalizes and improves the $O(lg \lg N)$ time solutions for the standard predecessor problem. Our data structure is the first with a query time that only depends on the numerical distance between the finger and the query integer.

Dynamic Partial Sums. The well-studied partial sums problem is to store a sequence of n integers with support for sum and search queries. The sequence is static in the sense that its length cannot change, but the update operation can be used to change the value of an integer in the sequence by a given value. There are matching lower and upper bounds showing that the problem can be solved on the w-bit Word RAM in linear space and $O(lg n = lg(w=))$ time per operation, where $w$ is the maximum number of bits allowed in updates.

As a natural generalization we consider dynamic partial sums, allowing insertions and deletions in the sequence. Our solution requires linear space and supports all operations in optimal worst-case time $O(lg n/ lg(w/δ))$, matching lower bounds for all supported operations. Our data structure is the first dynamic partial sums solution that matches the lower bounds, and the first to support storing integers of more than $lg w$ bits.