Spatio-temporal Range Searching Over Compressed Kinetic Sensor Data

[Abstract]

Sorelle A. Friedler*
sorelle@cs.umd.edu
http://www.cs.umd.edu/~sorelle

David M. Mount†
mount@cs.umd.edu
http://www.cs.umd.edu/~mount

Dept. of Computer Science, University of Maryland, College Park, MD 20742

Sensor networks and the data they collect have become increasingly prevalent and large. Sensor networks are frequently employed to observe objects in motion and are used to record traffic data, observe wildlife migration patterns, and observe motion from many other settings. In order to perform accurate statistical analyses of this data over arbitrary periods of time, the data must be faithfully recorded and stored. Large sensor networks, for example observing a city's traffic patterns, may generate gigabytes of data each day. The vast quantities of such data necessitate compression of the sensor observations, yet analysis of these observations is desirable. Ideally, such analysis should operate over the compressed data without decompressing it. Before more sophisticated statistical analyses of the data may be performed, retrieval queries must be supported. In this work, we present the first range searching queries operating over compressed kinetic sensor data.

In an earlier paper, we presented an algorithm for losslessly compressing kinetic sensor data and a framework for analyzing its performance. We assume that we are given a set of sensors, which are at fixed locations in a space of constant dimension. (Our results apply generally to metric spaces of constant doubling dimension.) These sensors monitor the movement of a number of kinetic objects. Each sensor monitors an associated region of space, and at regular time steps it records an occupancy count of the number of objects passing through its region. Over time, each sensor produces a string of occupancy counts, and the problem considered previously is how to compress all these strings.

Previous compression of sensor data in the literature has focused on stream algorithms and lossy compression of the data. We consider lossless compression. This is often more appropriate in scientific contexts, where analysis is performed after the data has been collected and accurate results are required. Lossless compression algorithms have been studied in the single-string setting, but remain mostly unstudied in a sensor-based setting.

In order to query observed sensor data, which ranges over time and space, we need to consider both temporal and spatial queries. *Temporal range queries* are given a time interval and return an aggregation of the observations over that interval. *Spatial range queries* are given some region of space (e.g., a rectangle, sphere, or halfplane) and return an aggregation of the observations within that region. *Spatio-temporal range queries* generalize these by returning an aggregation restricted by both a temporal and a spatial range. We assume that occupancy counts are taken from a commutative semigroup, and the result is a semigroup sum over the range. There are many different

*The work of Sorelle Friedler has been supported in part by the AT&T Labs Fellowship Program.

†The work of David Mount has been supported in part by the National Science Foundation under grant CCR-0635099 and the Office of Naval Research under grant N00014-08-1-1015

Bounds for Range Searching

	Temporal	Spatio-temporal
Preprocessing time	$O(\text{Enc}(X))$	$O(\text{Enc}(\mathbf{X}))$
Query time	$O(\log T)$	$O(((1/\varepsilon^{d-1}) + \log S) \log T)$
Space	$O(\text{Enc}(X))$	$O(\text{Enc}(\mathbf{X}) \log S)$

Table 1: Asymptotic time and space results achieved for temporal range searching and ε-approximate spherical spatio-temporal range searching in $\mathbb{R}^d$ space, where S is the number of sensors in the network, T is the length of the observation period, and $\text{Enc}(X)$ and $\text{Enc}(\mathbf{X})$ denote the size of the compressed representations of a single sensor stream (for temporal range searching) and sensor system (for spatio-temporal range searching).

data structures for range searching (on uncompressed data), depending on the properties of the underlying space, the nature of the ranges, properties of the semigroup, and whether approximation is allowed.

We present data structures for storing compressed sensor data and algorithms for performing spatio-temporal range queries over this data. We analyze the quality of these range searching algorithms in both time and space by considering the information content of the set of sensor outputs. There are two well-known ways in which to define the information content of a string, classical Shannon entropy and empirical entropy. Shannon entropy is defined in a statistical context, under the assumption that X is a stationary, ergodic random process. The normalized Shannon entropy, denoted $H(X)$, provides a lower bound on the number of bits needed to encode a character of X. In contrast, the empirical entropy, denoted $H_k(X)$, is similar in spirit to the Shannon entropy, but assumes no underlying random process and relies only on the observed string and the context of the most recent k characters.

Previous retrieval over compressed text (without relying on decompression) has been studied in the context of strings and XML files. For example, Ferragina and Manzini show that it is possible to retrieve all occurrences of a given pattern in the compressed text with query time equal to the number of occurrences plus the length of the pattern. Their space requirement is $5T \cdot H_k(X) + o(T)$ bits for a string X of length T. However, their data structure allows substring queries, which are very different from semigroup range searching queries, which we consider here.

Results

In this work we present the first range query results over compressed data; our results apply specifically to compressed kinetic sensor data. In addition we generalize the analysis of a previously presented compression algorithm to hold in an empirical context. We analyze the range query results in both a statistical and an empirical context. The preprocessing bounds show that we only make one pass over the compressed data. The query bounds are logarithmic in the input size. The space bounds, given in bits, show that we achieve these results without decompressing the data. For specific bounds see Table 1. Our temporal bounds rely on an extension of the Sleator and Tarjan data structure for dynamic trees to query time periods and aggregate underlying data. Our spatio-temporal bounds combine these temporal aggregations with a variant of approximate range searching that relies on the separation properties of the compressed data.