An Efficient Index Structure for Shift and Scale Invariant Search of Multi-attribute Time Sequences

Tamer Kahveci
Department of Computer Science

Ambuj Singh
Department of Computer Science

Aliekber Gürel
Department of Mathematics

University of California, Santa Barbara, CA 93106
{tamer,ambuj}@cs.ucsb.edu, aliekber@math.ucsb.edu

Time series or sequence data sets arise naturally in many real world applications like stock market, weather forecasts, video databases, sensor-based controls, and medicine. There is a frequent need to understand the information content of this data in order to respond better to common trends, to provide corrective emergency steps, or to predict the future evolution based on past records.

Defining the distance as some norm of the difference between two time sequences may be insufficient if the sequences can be made closer by linear transformations. The most important transformations are scaling and shifting. Scaling is needed because of the need to compare time sequences recorded on devices with different calibrations or different units. For example, a medical sensor in North America may record body temperature in $^\circ F$ while a similar sensor in Europe may use $^\circ C$. A sensor may also lose exact calibration over time (e.g. weather sensors on a satellite) leading to a need to compare measurements after different scalings. A similar case can be made for comparing time series data after eliminating shifts. Usually this is because the zero value in a time sequence differs in different measurements, and the results of a query need to be invariant under all possible shiftings.

In this paper [1], we considered the problem of shift and scale invariant search for multi-attribute time sequences. Our work fills a void in the existing literature for time sequence similarity since the existing techniques do not consider the general symmetric formulation of the problem. We define a new distance function for multi-attribute time sequences that is symmetric: the distance between two time sequences is defined to be the smallest Euclidean distance after scaling and shifting either one of the sequences to be as close to the other. We define two models for comparing multi-attribute time sequences: in the first model, the scalings and shiftings of the component sequences are dependent, and in the second model they are independent.

We propose a novel index structure called *CS-Index* (Cone Slice) for shift and scale invariant comparison of time sequences. As a part of this technique, the sequences in the database are first mapped to the shift eliminated plane. The transformed points are then clustered in hierarchical cone slices. These slices are stored on disk according to an *in-order* traversal, and a pointer to each slice along with angle and spatial extent information is maintained in memory. Given any query, it is first mapped to the shift eliminated plane. The shift and scale invariant distances between the query and the slices are computed in memory to obtain a set of candidate slices. The hierarchical construction of the index structure allows early pruning. Finally, the candidate slices are read from disk in a single seek, and false hits are eliminated. In-order storage of the slices on disk reduces the number of pages that need to be read. We show that the CS-Index structure can be extended easily to support multi-attribute sequences.

Experimental results show that the CS-Index structure performs 5 to 10 times faster than the R*-tree index structure and sequential scan. The efficiency of the index structure can be further improved by selectively replicating or caching parts of the index structure. If the pages at the first level of the index structure are replicated on disk, then the performance of our method improves further by a factor of two. Replicating one more level improves the performance slightly, but further replication has an adverse effect. Caching of the index structure (instead of replication) can also be used to improve the performance of the index structure. In our experiments, caching only one level of the index structure led to a performance improvement of 10 to 25 times over other techniques.

Based on the idea of using multiple resolution levels, we proposed another index structure called MRCS-index that clusters the subsequences for different resolutions in a grid of cone slices. The MRCS-index structure is a dynamic index structure and it allows subsequence search for queries of arbitrary length. According to our experiments, the MRCS-index structure performs 3 times faster than sequential scan.

References

[1] T. Kahveci, A.K. Singh, and A. Gurel. Shift and scale invariant search of multi-attribute time sequences. Technical report, UCSB, 2001.