Similarity Search in Streaming Time Series Based on MP_C Dimensionality Reduction Method

Thanh-Son Nguyen and Tuan-Anh Duong

1 Faculty of Information Technology, HCM City University of Technical Education
2 Faculty of Computer Science and Engineering, HCM City University of Technology

Abstract. The similarity search problem in streaming time series has become a hot research topic since such data arise in so many applications of various areas. In this problem, the fact that data streams are updated continuously as new data arrive in real time is a challenge due to expensive dimensionality reduction recomputation and index update costs. In this paper, adopting the same ideas of a delayed update policy and an incremental computation from IDC index (Incremental Discrete Fourier Transform (DFT) Computation – Index) we propose a new approach for similarity search in streaming time series by using MP_C as dimensionality reduction method with the support of Skyline index. Our experiments show that our proposed approach for similarity search in streaming time series is more efficient than the IDC-Index in terms of pruning power, normalized CPU cost and recomputation and update time.

1 Introduction

A streaming time series (STS) is a real-value sequence \( C = c_1, c_2, \ldots \), where new values are continuously added at the end of the sequence \( C \) as time progresses. Because a STS includes a great number of values, similarity is measured with the \( W \) last values of the streams (\( W \) is the length of a sliding window).

The similarity search problem in STS has become a hot research topic due to its importance in many applications of various areas such as earthquake forecast, internet traffic examination, moving object examination, financial market analysis, and anomaly detection ([2],[5],[6]). The challenge in these applications is that STS comes continuously in real time, i.e., time series data are frequently updated. Methods that are used to perform similarity search on archived time series in the past may not work efficiently in streaming scenarios because the update and re-computation costs in STS are significant. Therefore one needs an efficient and effective method for similarity search in this time series data type.

In [5] and [6], Kontaki et al. proposed an index structure, IDC-Index (Incremental DFT Computation – Index), which can be used for the similarity search in STS. The IDC-Index is based on a multi-dimensional index structure, R*-tree, improved with a delayed update policy and an incremental calculation of DFT. This approach is used in order to reduce update and re-computation costs in STS similarity search. However, the effect of using IDC-Index is not high due to using R*-tree as index structure and DFT as dimensionality reduction method. In order to enhance the efficiency of similarity search...
in STS we propose a new approach which uses MP_C, as dimensional reduction method and Skyline index as multidimensional index structure.

In the proposed approach, while using the same idea of a delayed update strategy and an incremental calculation for feature extraction, we can show that our new method for time series dimensionality reduction, MP_C, with the support of Skyline index can provide a more efficient similarity search in STS than IDC-Index in terms of pruning power, normalized CPU cost and recomputation and update time.

2 Preliminaries

2.1 Index Structures

The popular multidimensional index structures are R-tree and its variants ([1], [3]). In a multidimensional index structure (e.g., R-tree or R*-tree), each node is associated with a minimum bounding rectangle (MBR). A MBR at a node is the minimum bounding box of the MBRs of its child nodes. A potential weakness in the method using MBR is that MBRs in index nodes can overlap. Overlapping rectangles could have negative effect on the search performance. Besides, another problem in the method using MBR is that summarizing data in MBRs, the sequence nature of time series is not captured.

Skyline Index, another elegant paradigm for indexing time series data which uses another kind of minimum bounding regions, is proposed by Li et al., 2004 [9]. Skyline Index adopts new Skyline Bounding Regions (SBR) to approximate and represent a group of time series data according to their collective shape. An SBR is defined in the same time-value space where time series data are defined. Therefore, SBRs can capture the sequential nature of time series. SBRs allow us to define a distance function that tightly lower-bounds the distance between a query and a group of time series data. SBRs are free of internal overlaps. Hence using the same amount of space in an index node, SBR defined a better bounding region. For k-nearest-neighbor (KNN) queries, Skyline index approach can be coupled with some well-known dimensionality reduction technique such as APCA and improve its performance by up to a factor of 3 ([9]).

2.2 Similarity Search in Streaming Time Series

Many solution approaches have been proposed for similarity query in streaming time series lately. In [4] Gao and Wang, 2002, proposed a method based on a prediction for similarity search in STS. In this case, time series data are static and the query is changes over time. The authors solve the problem by using Fast Fourier Transform (FFT) to find out the cross correlations between the query and time series. The Euclidean distance between the query and each time series is calculated based on the predicted values. When the actual query is incoming the prediction error and the predicted distances are used to discard false alarms.

In [8], Liu et al., 2003, proposed a model for processing STS based on an index structure that can adapt to the change in the length of data objects. In this work, the VA-stream and VA+-stream index structures are used to query k-nearest neighbors. These methods partition the data space into $2^b$ cells, where $b$ is defined by user. They distribute different number of bits for each dimension so that the total of these bits is