Multi-Scale Histograms for Answering Queries over Time Series Data

Lei Chen and M. Tamer Özsu
University of Waterloo
School of Computer Science
{l6chen,tozsu}@uwaterloo.ca

Similarity-based time series data retrieval has been used in many real world applications, such as stock data or weather data analysis. Two types of queries on time series data are generally studied: pattern existence queries and exact match queries. In pattern existence queries, users are interested in the general shape of time series data and ignore the specific details. For example, users may want to retrieve all the stock data of last month that have a *head and shoulder pattern*. As long as the time series data have the specified pattern, they will be retrieved, no matter when the pattern appears and how it appears. For exact match queries, the exact result of a query is defined in terms of specific values. The actual match results are the time series data that are within a specific threshold of the exact result. For example, users may ask for all the stock data of last month that is similar to the IBM's stock data of last month.

In this paper (see [1] for a full version), we describe a technique to answer both pattern existence queries and exact match queries. A typical application that needs answers to both queries is an interactive analysis of time series data. For example, users may be initially interested in retrieving all the time series data that have some specific patterns that can be quickly answered by pattern existence queries. After that, they may want to retrieve all the time series which are similar to an interesting time series that they find from the previous retrieval results. In this case, they can use exact match queries.

We propose a histogram-based representation to approximate time series data. Time series histograms are computed from normal form of time series data, the distance that is computed from two time series histograms are invariant to amplitude scaling and shifting. Furthermore, because time series histograms ignore the temporal information, they are also robust to time shifting and scaling. Moreover, since time series histograms show the whole distribution of the data and noise only make up a very small portion, comparisons based on histograms can remove the disturbance caused by noise. Therefore, time series histograms are ideal representations for answering pattern existence queries. The time series histograms give a global view of the data distribution of time series data. However, they do not consider the temporal appearance order of values. Therefore, we propose a multi-scale representation of time series histogram for better discrimination of time series data based on their temporal details to facilitate exact match queries. With multi-scale time series histograms, the exact match queries can be answered at several precision levels. Users can specify the scale levels when they submit a query. We also investigate two different approaches to construct histograms: *equal bin size* and *equal area size*. Weighted Euclidean distance is used to measure the similarity between two time series. For both pattern existence queries and exact match queries, directly comparing time series histograms is computationally expensive, therefore, we use multi-step filtering and the weighted average of time series histogram to avoid comparisons on full histograms.

Experimental results show that, compared to symbolic representation, our time series histogram-based approach can achieve relatively high precision and recall in answering pattern existence queries. Furthermore, with the proper setting of scales and number of bins, we observe that weighted Euclidean distances computed from multi-scale time series histograms outperform dynamic time warping and longest common subsequences in answering exact match queries when the time series data contain local time shifts and noise. The experiments also show that equal area size histograms are more suitable for time series data comparison and weighted average histogram distances can effectively prune the false alarms from the database before computing the weighted Euclidean distance between two histograms, which will be helpful when the database size becomes large.

Acknowledgements: Thanks to Eamonn Keogh, Michalis Vlachos, Changzhou Wang, and Xiaoyan Sean Wang, for providing their source codes or data sets.

References

[1] L. Chen and M. T. Özsu. Similarity-based retrieval of time-series data using multi-scale histograms. Technical Report CS2003-31, University of Waterloo, 2003.

Proceedings of the 20th International Conference on Data Engineering (ICDE'04)
1063-6382/04 $ 20.00 © 2004 IEEE