University of Zurich^UZH

Department of Informatics

University of Zürich
Department of Informatics
Binzmühlestr. 14
CH-8050 Zürich
Phone. +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch/dbtg

UZH, Dept. of Informatics, Binzmühlestr. 14, CH-8050 Zürich

Prof. Dr. Michael Böhlen
Professor
Phone +41 44 635 43 33
Fax +41 44 635 68 09
boehlen@ifi.uzh.ch

Zürich, September 29, 2020

MSc Project: Pair-wise Correlations Analysis among Multiple Time Series Data

A time series is simply a series of data points ordered in time. Time series analysis includes methods for analyzing time series data in order to extract meaningful statistics and characteristics of the data. In addition to single series statistics such as average and standard deviation, we also want to identify the strength of correlation among all pairs of series.

Correlation measures the relationship between two series. Intuitively, it indicates how changes in one series are associated with changes in a second series, *i.e.*, whether they are moving in the same direction or not. In this project, our main goal is to compute pair-wise correlations in a dataset consist of multiple time series. Recent works [1], [2], [3] have shown the importance of computing Pearson correlation of a large number of signals. Based on that, there are five important queries in a data center management application, out of which three queries related to server dependency analysis, load balancing, and anomaly detection involve computing correlation matrices.

However, it is not as simple as computing the correlation, as it would be extremely challenging for a large data set. For example, a warehousing system monitors multiple data centers, each of which may contain tens of thousands of servers. Assume that 500 performance counters are collected from each server. Then, data centers with 100,000 servers will yield 50 million concurrent time series and, with a mere 15-second sampling rate, more than 30 billion records (or about 1TB data) a day. The sheer volume of the data can make useful data mining queries, like computing correlation, impractically slow.

To address the problem, it makes sense to approximate time series by reducing its number of data points. This can be achieved by dimensionality reduction techniques such as Discrete Fourier transform (DFT) and Piecewise Aggregate Approximation (PAA). For many real applications, the number of highly correlated pairs is much smaller than the number of all possible pairs of series. On the other hand, users are typically interested in correlated pairs, and the uncorrelated ones are usually not of much interest. Therefore, while the exact correlations of correlated pairs need to be computed, that of uncorrelated pairs can be safely ignored using the approximation algorithms. The approximate solutions are as useful as corresponding exact solutions, however, they are much faster than the exact solutions.

The project is structured into the following tasks:

- The first step is doing some preparatory work in the form of exploratory data analysis on a time series dataset to better understand what time series data looks like and what cross correlation means for time series data.
- The first (main) task of the project is to implement several algorithms for computing the Pearson cross correlation for time series data. These algorithms fall into two groups: exact algorithms and approximation algorithms.
  - There are two exact algorithms: a naive implementation of the formula for the Pearson cross correlation and a faster incremental one.
  - There are three approximation algorithms, all based on a dimensionality reduction technique: Discrete Fourier transform (DFT), Piecewise Aggregate Approximation (PAA) and Discrete Wavelet Transform (DWT). Implementation of the algorithm based on the DWT is optional and will be considered as a bonus task.
- The second task is to develop an (automated) benchmark environment with which the algorithms can be evaluated experimentally. This includes tools to schedule and run experiments and to evaluate, process, and visualize the results of the experiments. This may also include tools to preprocess time series data to make it suitable as input for the algorithms.
- The third task is to systematically evaluate the performance of the different algorithms to check their suitability for certain application scenarios. Practically speaking, this means running the algorithms for different datasets under different parameter sets and to use the tools developed in the second task to visualize and present the results.

Supervisor: Sven Helmer, AmirReza Alizade Nikoo

Department of Informatics, University of Zurich

Prof. Dr. Michael Böhlen

References

[1] A. Mueen, S. Nath, and J. Liu. Fast approximate correlation for massive time-series data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 171–182, January 2010.

[2] G. Reeves, J. Liu, S. Nath, and F. Zhao. Managing massive time series streams with multi-scale compressed trickles. *Proc. VLDB Endow.*, 2(1):97–108, August 2009.

[3] Y. Zhu and D. E. Shasha. Statstream: Statistical monitoring of thousands of data streams in real time. In *Proceedings of 28th International Conference on Very Large Data Bases, VLDB 2002, Hong Kong, August 20-23, 2002*, pages 358–369, August 2002.