eXascale Infolab

Department of Informatics

University of Fribourg
Department of Informatics
Bd. de Pérolles 90
CH-1700 Fribourg
Phone. +41 26 300 84 65
Fax +41 26 300 97 26
http://exascale.info/

University of Fribourg, Dept. of Informatics, CH-1700 Fribourg

Adrian Hänni
Bern

Switzerland

Djellel Difallah/Mourad Khayati
Senior researchers
Phone +41 26 300 84 60
Fax +41 26 300 97 26
{ded,mkhayati}@exascale.info

Fribourg, 27. Juni 2016

Correlation-based LISA anomaly detection for time series

MSc Thesis:

Work overview:

Local indicators of spatial association (LISA) has been introduced as a measure of local clustering of similar measurements [1]. Consider a set of m time series $\mathcal{X} = \{X_1, X_2, \ldots, X_m\}$ where the p-th time series is a set of n observations and defined as $X_p = \{(t_{p1}, v_{p1}), (t_{p2}, v_{p2}) \cdots, (t_{pn}, v_{pn})\}$ with t_i as the timestamp of an observation and v_i as the value of that observation. The LISA of a value v_{pi} is defined as follows

$$L(v_{pi}) = z_{ip} \times \sum_{q=1}^{k \leq m} w_{pq} z_{qi} \tag{1}$$

where $z_{pi} = \frac{v_{pi} - \bar{v}_i}{\sigma_i}$ with $\bar{v}_i$, σ_i and k being respectively the mean, the standard deviation and the number of all v_{*i} values, and w_{pq} is a similarity weight between two time series X_p and X_q. The value of L is interpreted as follows:

- $L(v_{pi}) > 0$ indicates the presence of clusters of high or low values, or outliers conversely.
- The higher $L(v_{pi})$ is, the more dissimilar is v_{pi} compared to its k neighbors.

In [2], the LISA computation was applied to detect anomalies in time series data. For instance, the temporal dimension was added to LISA computation by extending the local neighborhood to contain both temporal and spatial neighbors. More specifically, the set of k neighbors of a value v_{pi} was extended to include its previous observations in addition to both the current-and-past observations from its neighbors. This solution yields accurate anomaly detection in Water Distribution Networks (WDNs), where the time series created by the sensors placed on directly connected pipes are assumed to be similar.

The goal of this thesis is to implement a correlation based LISA solution for anomaly detection in time series data. The implemented solution aims to generalize the solution proposed in [2] by automatically computing the correlation across the different time series produced by the sensors placed on the network. The implemented solution will identify and leverage the correlation that might exist between two time-series regardless of the position of the sensors and the type of the observation, e.g., temperature, pressure, etc.

Work tasks:

1. Familiarize yourself with LISA measure.
2. Implement LISA using correlation (pre-processing, incremental, moving window, etc.)
3. Implement a graphical tool to visualize the result of LISA-based anomaly detection on time series.
4. Write a thesis that describes the algorithm and the tool.
5. Presentation of 20 minutes.

Literature:

1. Luc Anselin: *Local indicators of spatial association—LISA*, in Geographical analysis 27.2 (1995): 93-115.
2. Djellel Eddine Difallah, Philippe Cudre-Mauroux, and Sean A. McKenna: *Scalable anomaly detection for smart city infrastructure networks*, in Internet Computing, IEEE 17.6 (2013): 39-47.
3. Spiros Papadimitriou, Jimeng Sun , and Philip S. Yu: *Local Correlation Tracking in Time Series* in Proceedings of the Sixth International Conference on Data Mining (ICDM), 2006.
4. Johannes Gehrke, Flip Korn, and Divesh Srivastava: *On Computing Correlated Aggregates Over Continual Data Streams*, in Proceedings of the ACM International Conference on Management of Data (SIGMOD), 2001

Starting date of thesis: 01.07.2016

Ending date of thesis: 01.02.2017

University of Fribourg
Department of Informatics
Exascale Infolab

Djellel Difallah & Mourad Khayati