

Exploration and Mining of Temporal Data

Ramoza Ahsan

PhD Candidate

WPI – Computer Science

Friday, December 8, 2017

Time: 9:00 a.m. – 10:00 a.m.

Location: Fuller Labs 141

Committee members:

Prof Elke A. Rundensteiner, Advisor, WPI - Computer Science

Prof Gabor Sarkozy, Co-advisor, WPI - Computer Science

Prof Xiangnan Kong, WPI - Computer Science

Prof Vassilis Athitsos, The University of Texas at Arlington - External member

Abstract:

Given the ubiquity of temporal data and the exponentially growth of datasets I propose to explore several research challenges in the area of mining such temporal data.

First, I address the problem of identifying frequent patterns in dynamic temporal data. Frequent pattern mining is computationally expensive, is parametrized and generates frequent patterns from scratch for every data update and different parameter settings. I propose an incremental method of generating a store of frequent patterns and updating in parameter space to assure near instantaneous retrieval of patterns. Efficient archival strategies to store, access and update these patterns over temporal data are also developed.

Second, I investigate the problem of finding similarities in temporal data using dynamic time warping distance. This powerful distance allows the flexible comparison of data of different lengths and temporal misalignments, however with prohibitively high computation costs. To achieve real time responsiveness, data reduction strategies are designed using inexpensive Euclidean distance with subsequent time warped matching on the reduced data.

Third, to explore correlations among time series data, I propose a framework that establishes a mapping between the popular non-metric Pearson correlations with the metric Euclidean distance. The model compresses the raw time series into Euclidean-based clusters augmented by a compact overlay graph encoding correlation relationships to support a rich diversity of correlation operations.

Fourth, extend the correlation analytics to work on large scale evolving data using distributed framework by introducing efficient strategies for data shuffling among worker nodes and pruning unnecessary correlation computations.

Comprehensive experimental studies will be conducted using real-world datasets as well as case studies to evaluate the efficiency and effectiveness of the above-proposed approaches compared to state-of-the-art methods.