University of Zurich[UZH]

Department of Informatics

University of Zurich
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 Zurich

Jonathan Nagel

Switzerland

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

Zürich, 14. März 2013

Centroid Decomposition Based Recovery for Segmented Time Series

BSc/MSc Thesis:

Work overview:

The Centroid Decomposition (CD) is a matrix decomposition technique that decomposes an $n \times m$ matrix $\mathbf{X} = [X_1|\ldots|X_m]$ into an $n \times m$ *loading* matrix $\mathbf{L} = [L_1|\ldots|L_m]$ and an $m \times m$ *relevance* matrix $\mathbf{R} = [R_1|\ldots|R_m]$ as follows:

$$CD(\mathbf{X}) = \mathbf{L}, \mathbf{R} \quad (1)$$

$$s.t. \quad \mathbf{X} = \mathbf{L} \times \mathbf{R}^T = \sum_{i=1}^{d} L_i \times R_i^T \quad (2)$$

Where $d \leq m$ is the number of dimensions to compute.

In this work, we use real world datasets that describe hydrological phenomena of up to 15 million observations produced by sensors in 242 mountain stations. In fact, Our hydrological database contains 79 temperature time series, 69 precipitation time series, 48 water level time series, 15 humidity time series, 4 wind speed time series and 3 air pressure time series. Due the snow met the sensors stop working and produce missing observations in the time series with the following percentages: 20% in temperature, 16% in precipitation, 66% in water level, 6% in humidity, 23% in wind speed and 11% in air pressure.

The aim of this thesis to investigate and implement the Centroid Decomposition method to recover blocks of missing values in time series. The special focus will be to compare the accuracy of the recovery technique using complete hydrological time series and segmented hydrological time series. The segmentation of the time series will be an issue to investigate.

Work tasks:

1. Understand and implement the Centroid Decomposition algorithm
2. Apply the Centroid Decomposition for the block recovery in hydrological time series. The datasets are already loaded in an Oracle server.
3. Review the time series segmentation techniques proposed in the literature
4. Implement a segmentation technique for the hydrological time series
5. Empirical comparison of the recovery accuracy of the implemented algorithm for complete and segmented time series
6. Report that describes the major parts of the work and a presentation in front of the Database Technology Group (15 min)

Literature:

1. Chu, M.T., and Funderlic, R.E., : *The Centroid Decomposition: Relationships Between Discrete Variational Decompositions and SVDs*, in SIAM J. Matrix Analysis and Applications, 2002
2. Börzsönyi., E., *Recovery of Missing Values based on Centroid decomposition*, Master Project, 2013
3. Zhang, R., Kotagiri, R. and Parampalli, U., *An Adaptive Algorithm for Online Time Series Segmentation with Error Bound Guarantee*, EDBT, 2012
4. Keogh, E., Chu, S., Hart, D., Pazzani, M., *An Online Algorithm for Segmenting Time Series*, ICDM, 2001
5. Benjamin, M.M., *Missing Data Problems in Machine Learning*, PhD thesis, 2008

Task assignment and supervisor:

- Mourad Khayati (mkhayati@ifi.uzh.ch)

Starting date of thesis: 05.03.2013

Ending date of thesis: 04.09.2013

University of Zurich
Department of Informatics

Prof. Dr. Michael Böhlen
Professor