**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, 15. Dezember 2015

**MSc Basismodul: Implementing the Case Matching Similarity**

A time series $s$ is a sequence of data points collected over a (possibly infinite) time interval and is frequently stored in a time series database $DB = \{s_1, s_2, \ldots, s_n\}$. Such data appears in many applications, e.g., the financial stock market, meteorology, sensor networks and network monitoring to name only a few. Time series data is often incomplete, though, due to sensor failures or other issues. In our research we developed the Top-$k$ Case Matching (TKCM) algorithm to *impute* missing values in meteorological time series, that is to replace the missing values with good estimates of what the values could have been.

TKCM exploits the fact that time series are often highly correlated. For example, a temperature time series recorded in Zurich is highly correlated to one recorded in Bern, but less correlated to one recorded in New York. TKCM defines for each (base) time series a set of reference time series. To impute a missing value at time series $s$, TKCM looks at the current values of the reference time series of $s$ and looks for past situations where the values of the reference time series were similar. It then derives the missing value from those past similar situations.

To choose the set of reference time series we developed the Case Matching Similarity (CMS) algorithm, designed to exploit the properties of TKCM. While the Pearson Correlation Coefficient (PCC) [1], another popular correlation function, measures the linear correlation between two time series, CMS measures the strength of co-occurrence between values of $s_1$ and $s_2$.

CMS splits the *range* of time series $s_1$ into equal-sized sub-ranges, called buckets. Each bucket contains the values of $s_2(t)$ such that $s_1(t)$ is within the bucket limits. Then CMS computes the amount of variation within each bucket. The smaller the variation is, the more similar are the values $s_2(t)$ in a bucket and hence the stronger is the co-occurrence.

Let $w \in \mathbb{R}_{>0}$ denote the bucket width and $z \in \mathbb{Z}$ the ID of a bucket. Then a bucket $b_z$ a list of the values $s_2(t)$ such that $s_1(t)$ is in the range $[zw, (z+1)w)$. More formally,

$$b_z = \{s_2(t) \mid \forall t : zw \leq s_1(t) < (z+1)w\}$$

Moreover, we define the bucket mean $\bar{b}_z$ and the bucket standard deviation $\sigma_z$ as follows

$$\bar{b}_z = \frac{1}{|b_z|} \sum_{s_2(t) \in b_z} s_2(t) \qquad \sigma_z = \sqrt{\frac{1}{|b_z|} \sum_{s_2(t) \in b_z} (s_2(t) - \bar{b}_z)}$$

Each bucket $b_z$ has a certain standard deviation $\sigma_z$, which is a measure of variation (or dispersion) inside the bucket. The smaller $\sigma_z$ is, the closer are the values $s_2(t) \in b_z$ to the bucket mean $\bar{b}_z$.

Let $B = \{b_z \mid \forall z \in \mathbb{Z} : b_z \neq \emptyset\}$ be the set of all non-empty buckets. Notice that the buckets in $B$ cover the entire range of $s_1(t)$. Then CMS is defined as the average bucket standard deviation, where each term is weighted by the number of elements in the corresponding bucket.

$$\text{CMS}(s_1, s_2) = \frac{1}{|B|} \sum_{b_z \in B} \frac{|b_z|}{|s_1|} \sigma_z$$

## Tasks

- Study, understand and implement PCC [1] and CMS.
- Analyze the runtime complexity and memory consumption of CMS.
- Summarize your findings in a short report.

## Optional Tasks

- Implement CMS$^+$ [2], which improves the runtime and space complexity of CMS by incrementally computing the standard deviation $\sigma_z$.
- Analyze the runtime and space complexity of CMS$^+$.

## References

[1] A. Mueen, S. Nath, and J. Liu. Fast approximate correlation for massive time-series data. In *SIGMOD'10*, pages 171–182, New York, NY, USA, 2010. ACM.

[2] K. Wellenzohn, M. Böhlen, A. Dignös, J. Gamper, and H. Mitterer. Imputation of missing values in highly correlated streams of time series data. Unpublished, 2015.

**Supervisor:** Kevin Wellenzohn (wellenzohn@ifi.uzh.ch)

University of Zurich
Department of Informatics

Prof. Dr. Michael Böhlen
Professor