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, 27. Oktober 2015

MSc Basismodul: Similarity Search in a Time Series Database

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 become often very large; sensors in jet engines collect, for example, time series data in the order of terabytes *each day*. Therefore algorithms to efficiently process large amounts of time series data are needed.

Some of the most frequent questions analysts ask about time series data are concerned with the *similarity* of two time series, e.g. *what was the most similar summer to this year's summer?* More specifically, given some time series s, one wants to find the time series $s' \in DB$ that is most similar to s according to some similarity function.

The Euclidean Distance (ED) is one of the most popular ways to quantify the similarity of two time series. Let $|s|$ be the length of time series s and $s(t)$ denote the value of s at time t then the Euclidean Distance is defined as:

$$ED(s_1, s_2) = \sqrt{\sum_{t=1}^{|s_1|}(s_1(t) - s_2(t))^2}$$

Another popular similarity measure is Dynamic Time Warping (DTW) that was initially developed to quantify the similarity of spoken words in speech recognition [3]. The algorithm "warps" the time axis to more accurately compute the similarity of two words spoken at different speed. Much effort has gone into reducing DTW's complexity to make it suitable for large scale time series processing [2, 4].

The aim of this project is to familiarize with the two algorithms and implement them. The student is asked to download the UCR Time Series Classification Archive [1] and build a small program that allows the user to query for the most similar time series in the archive.

Tasks

- Get to know the data set [1].
- Study, understand and implement both ED and DTW.
- Analyze the runtime complexity and memory consumption of both algorithms.
- Optimize the runtime of DTW using the Sakoe-Chiba Band [4].
- Summarize your findings in a short report.

Optional Tasks

- Optimize the memory consumption of DTW. (**Hint:** Do you need *all* the information in the matrix used for dynamic programming?)
- Optimize the runtime of DTW using the Itakura Parallelogram [4].

References

[1] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The ucr time series classification archive, July 2015. `www.cs.ucr.edu/~eamonn/time_series_data/`.

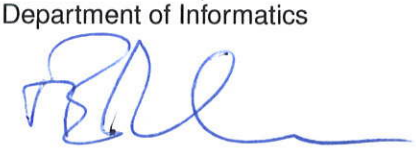
[2] E. Keogh. Exact indexing of dynamic time warping. In *Proceedings of the 28th International Conference on Very Large Data Bases*, VLDB '02, pages 406–417. VLDB Endowment, 2002.

[3] J. B. Kruskal and M. Liberman. The symmetric time-warping problem: from continuous to discrete. In D. Sankoff and J. B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules - The Theory and Practice of Sequence Comparison*, chapter 4. CSLI Publications, Stanford, CA 94305, 1983.

[4] S. Salvador and P. Chan. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.*, 11(5):561–580, Oct. 2007.

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

University of Zurich
Department of Informatics



Prof. Dr. Michael Böhlen
Professor