Mining of Process Data using User Defined Curve Patterns

T. Barz, O. Frey, J. Huss, L. Urbas
Technische Universitaet Berlin, Zentrum Mensch-Maschine-Systeme, Jebensstrasse 1, 10623 Berlin, Germany;
tel: ++49 (0)30 31429640, barz@zmms.tu-berlin.de, fax: ++49 (0)30 31472581

Over the last years storage of process data in chemical industry has turned out to be a huge database of widely unused information. This situation is comparable in some respect to the situation of the internet before the invention of search engines. Whereas the WWW search engines often use keywords for matching with literary language our approach is aimed at the match of trends in recorded signals. Variation in time of sensor data is commonly used by engineers to get an overview of process dynamics and history. While reading these curves, experts form implicit mental descriptions of the relationship between features and process conditions. Therefore we can use the engineer's process knowledge and their skills in curve reading for computerised process data exploration. Those descriptions can be explained by segmentation and categorisation of time series into a sequence of symbols, here called trendlets.

The interaction with a tool for the identification of specific process states and operating conditions should account for the user's comprehension of temporal characteristics in time series. Furthermore the algorithm should be based on the declarative representation of process trends and use this patterns for the search in historical process data. With the goal to provide users assistance in exploration of historical data for analysis, diagnosis and decision making in operation of chemical processes we present a prototype which covers both requirements. Our approach uses an algorithm which extracts and analyses features (trendlets) corresponding to the operators view of signal progression. Furthermore the method doesn't assume any form of model and relies only on process data. This facilitates an intuitive handling and interpretation of the applied method.

Transformation of time series into feature space is carried out by Continuous Wavelet Transformation (CWT) and Wavelet Transform Modulus Maxima (WTMM). This means time translation invariance of the result is guaranteed which is crucial for pattern identification. Efficient pre-processing by fast convolution of the time series covers the range of frequency bands in which features are to be expected. The Wavelet used for CWT is the second derivative of the gauss curve, the so called Mexican Hat Wavelet. Gaussian filtering guarantees optimal time-frequency resolution of features in the original signal. The resulting coefficients represent the curvature on the corresponding scales. The identification of significant trends in the signal is realized by extraction of the local modulus maxima on each scale. This is based on application of multiscale representation of process data via interval trees over the scale space. Stable episodes of maxima coefficients over the scale space represent meaningful features in the original data. Isolation of these maxima lines and neglect of the remaining coefficients preserves trend information completely and reduces data efficiently. This allows minimal data storage requirements and optimization of search speed through elimination of redundant information. The extracted features correspond directly to the trendlets in the selected time intervals defined by the user's query. The query consists in segments and corresponding trendlets and is transformed in the same way as the historical process data. It explicitly expresses the users understanding of the feature and frequency space he is interested in. The applied method allows performing similarity measures directly in the feature space. Thus re-transformation of the signal is not needed. Time and amplitude

migration of the maxima in the neighborhood of the selected scale represent the trend. Query and historical data are compared using a weighted L^{∞}-Norm as distance measure to approximate the Hölder exponent on the given scales of Gaussian smoothening, instead of the signal itself. For linear segments additional information about the first order derivative is used.

The principles for the interaction with the graphical interface and the robustness of search will be presented by application to experimental data of a coupled column system. The capability of the prototype for users assistance will be demonstrated by exploring specific plant behaviour in historical data. The provision of additional information as support for diagnosis and decision making trough start up of the high pressure column and search for temporal violation of operational restrictions will be shown. Finally performance measures like requirements for data storage and computational efficiency are discussed.