54 KNOWLEDGE ACQUISITION FROM HISTORICAL DATA FOR CASE ORIENTED SUPERVISORY CONTROL

Alexei Lisounkin*, Gerhard Schreck*, Hans-Werner Schmidt**

(*) Fraunhofer-Institute for Production Systems and Design Technology
Pascalstraße 8-9, D-10587 Berlin, GERMANY
e-mail: {alexei.lisounkin, gerhard.schreck}@ipk.fhg.de

(**) ELPRO Prozessindustrie- und Energieanlagen GmbH
Marzahner Straße 34, D-13053 Berlin, GERMANY
e-mail: hans-werner.schmidt@elpro.de

This paper presents a knowledge acquisition procedure based on data mining methods and its integration within a SCADA environment for decision support of plant operators. The results shown are part of investigations using real historical data of water treatment plants.

1. INTRODUCTION

Even for high level automation of process supervision, diagnostics, and control, the facility operator role is continually increasing. Although local automation tasks are covered by an installed control system, the facility operation staff take precaution with high level functional, technological, strategic objectives. Here, human experience plays an unique role. The aim of the knowledge-based methods is to connect objective information from the facility – available historical trends and data – with experience-based evaluation and assessment through an operator team.

Use of data mining approach will support elaboration of case characteristic information and ensure objectivity of the decision making procedure. The elaboration of a generic data mining approach for the process control knowledge acquisition is the focus of this paper. The corresponding functional sequence was developed and applied for the implementation of high level control and supervision tasks for water treatment plants.

The main aim of the procedure is to extract valuable information – knowledge – from a historical data series. The analysis of the water consumption profiles is subordinate to the middle-term and long-term facility control tasks, as well as simulation based operator training. The following chapters give a description of the relevant steps and give numerical examples as illustration.

2. CASE ORIENTED SUPERVISORY CONTROL

The increasing complexity of modern process plant and the demands for energy conservation, product quality, environment protection, safety and reliability make new approaches to process automation necessary. Beside decision-making and control procedures based on mathematical models, and solvers for multi-criteria optimization tasks, knowledge based systems involving the experience of human operators are a promising approach.

SCADA systems (supervisory control and data acquisition) are usually introduced for high level process control and automation. They play an important integration function in distributed, multilevel control environments, and provide the common view to the system and according operator panels required for process operation. Typical functionalities include activation of control actions, monitoring of process states, recording of alarms and events, emergency shutdown, etc.

The human operator team is responsible for the high level process management which includes tasks like

- Assessment of process situations,
- Selection of operating points,
- Consideration of different modes of working and use of resources,
- Reaction to changing requirements / demands,
- To ensure a continuous and smooth running of the system.

Operator decisions are based on its knowledge on process situations, process trends, and process control. Hereby the experience gained by the past – historical data & situations – plays an important role to reach high level performance. Case oriented supervisory control means a mapping of known process situations to approved control strategies & actions. Therefore the identification of operation profiles of processes or sub-processes can be identified as a basic function of a decision support system. Figure 1 presents the basic concept of decision support and knowledge acquisition based on data mining. It considers the long term development of a knowledge base on approved control patterns as well as the short term decision support on actual process situations. This results to an adaptive system with high flexibility and active involvement of the human operators.

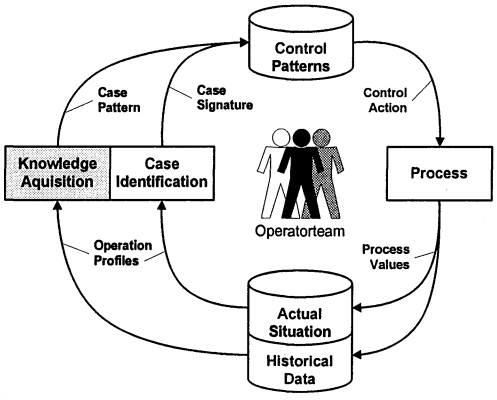

Figure 1 – Knowledge acquisition from operation profiles.

3. KNOWLEDGE ACQUISITION PROCEDURE

3.1 General Aspects

Traditionally, data analysis techniques consist of a sequence of data processing algorithms which investigate general characteristics of data series – such as minimum maximum, and mean values, statistics, images with respect to a given convolution operator, etc. – and then leave semantic data interpretation to a human. The data exploitation aspects – pragmatics – impact semantics of the entire data analysis procedure dramatically, which is reflected in the choice and parameterization of data processing algorithms. For this reason, the data analysis procedure is considered to be data driven knowledge acquisition, and our objective is to emphasize the semantic aspects for all steps of the data processing chain.

The principle steps of the data driven knowledge acquisition chain are summarized in the following list:

- data acquisition,
- data conditioning (validation, regularization and pre-processing),
- definition of semantics,
- knowledge extraction.

In Figure 2, the steps of the knowledge acquisition chain are associated with corresponding general data processing methods. This scheme defines a framework for configuration of the chain with respect to semantics and operational aspects.

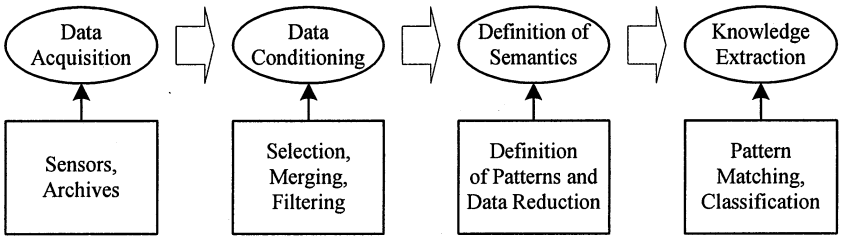

Figure 2 – Principle chain of data driven knowledge acquisition.

Special attention must also be paid to the characterization of the time enlargement of the process. Here, a continuous data flow must be separated into characteristic units possessing, in some sense, common information features. Thus, identification of repeatable operating sequences plays an important role with respect to the exploitation of data analysis results for case oriented process supervision and control.

Further, the process profile denotes a set of chronologically ordered data samples enlarged within the specified time interval. For the investigation of water consumption profiles, we set the profile time enlargement to 24 hours. This time interval corresponds to the sleep-wake living cycle of people and, moreover, it is explicitly recognizable in the process data flow.

3.2 Data Acquisition

Data samples consist of practical measurements, status information, and control signals available for the analyzed process – so-called process values. In cases where a water supply and distribution facility is controlled by a SCADA system, these process values are supplied by instruments installed in the physical system and by the logic of the control components itself. The SCADA system is usually equipped with a database which saves the process values. The archiving can be event-oriented – when a certain condition is being met – or raster-oriented – cyclic with respect to a defined time period.

For example, in Figure 3a, a single water consumption trend for a small German city (ca. 200,000 inhabitants) is depicted. Such trends collected over the period of two years have been analyzed by the authors and have offered the required input for the investigations.

Additionally, new data items within the data sample can be generated with an algorithmic analysis of the measured data sample components. Such a procedure is known as data transformation. The linear and non-linear composition of components, numerical integration and derivation are examples of data transformation. Complimentary to the transformation, data selection should also be mentioned at this point. It is reasonable to use only a subset of process values for further consideration. The other archived data items (components) will be omitted as irrelevant.

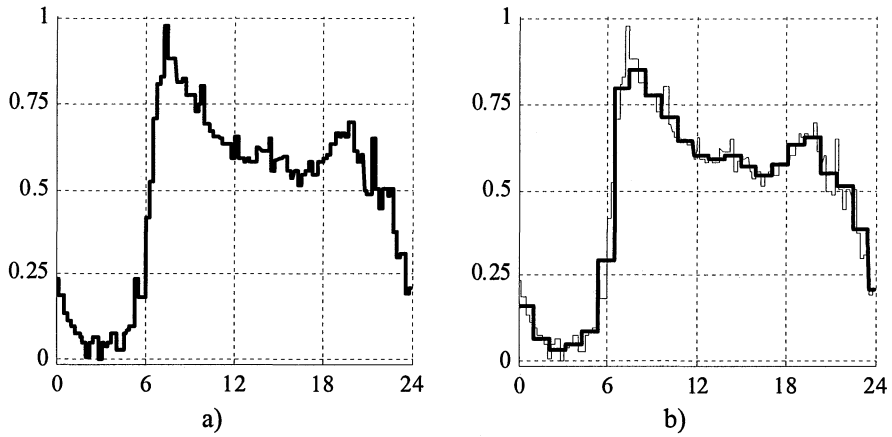

Figure 3 – a) Water consumption profile (abscise – time in hour, ordinate normalized); b) Profile after data filtering.

3.3 Data Conditioning

The objective of this step of the data analysis has is to validate data profiles and standardize their time raster, as well as to change some data properties (e.g., disturbances and noise reduction). In the beginning, the practical data profiles must be checked with respect to gaps and inaccurate data. The data validation is based on intuitive knowledge of the process – possible minimal/maximal values, expected

data sampling time, etc. Inaccurate data can be set to some neutral values, gaps in the data time series can be closed by interpolation or by involving statistical information. An alternative option is to exclude profiles with gaps or inaccurate data from being considered further.

The standardization of data means the obtaining of a common time lattice (common sampling cycle) for all considered data profiles. The common time lattice is important for the profile classification procedures which operate on sets of profiles by converting them into matrix form.

Any measurement is a raw noisy data, which, in a lot of cases, must be prepared for the next steps of data analysis. Traditionally, noise reduction is proceeded by regression filters, such as low-pass or band-pass filters. The regression filters are profitable for cases in which the regression model is adequate for forecasting the process dynamics. In the described application domain, we can not expect an adequate modeling from the regression. Moreover, the use of regression filters for smoothing out the noise has the disadvantage of stretching and flattening the data: valleys and peaks become wider and their magnitude becomes smaller.

Recently, wavelet based filters have extensively been used for data processing. Here, the multi-resolution analysis is applied for the identification of meaningful data and to "smooth" out the measurement noise (Lisounkin, 2003). The shape of the mother-wavelet must be chosen in accordance with profile interpreting and archiving pragmatics. In Figure 3b, the result of water consumption profile filtering by means of the wavelet filter is depicted. Here, the Haar-mother-wavelet was selected in order to ensure minimal number of system switches and steady-state conditions in the intervals between the switch points.

3.4 Definition of Data Semantic

The objective of the data semantic analysis is to elaborate such data characteristics and a criterion which allow the interpretation of actual process data with respect to data exploitation procedure. Rules for the interpretation of the information hidden in data profiles must be defined by a human. Simultaneously, process information which is declared unimportant, will be omitted. The definition of data semantic is indeed the process of data abstraction.

An abstraction of the data is usually connected with the definition of data semantics. Thus, a semantic sensible reduction of data is applied in order to emphasize relevant features of the profile and to substitute the non-relevant information with neutral values with respect to the comparison procedure. This procedure can be characterized as morphologic processing of the data. The morphologic processing completely changes the nature of the data and represents a semantically conditioned data reduction.

With respect to water facilities supervision and control, such data which represents dynamic behavior of the water consumer is highly instructive for facility operation. For this reason, high and low amounts of water consumption were investigated, and a procedure for water consumption forecasting was developed. From this point of view, cases of high and low water consumption in water profiles are the semantic payload of the data. For the analysis and forecasting of high and low water consumption, the data model based on profile string coding and approximating string matching approach has been applied.

3.5 Extraction of Knowledge

Knowledge extraction from data series can be characterized by one of the following approaches (Müller, 2000):

- deviation detection and change measure – discovering the most significant changes in the data from previously measured or normative values,
- clustering – identifying a finite set of categories that describe the data,
- regression analysis – learning a function that maps a data item in a prediction variable,
- dependencies modeling – finding a model that describes significant dependencies between variables.

For the middle-term and long-term facility simulation and control tasks, the clustering approach has been considered. Here, the main objective of the knowledge extraction is identification of representative process data profiles and mapping of them onto the process and its context characterizations. The guides for the mapping procedure must be provided by facility staff.

It was assumed that the basis for the data driven knowledge extraction should be a set of data profiles, which possess high versatility with respect to possible facility (process) conditions.

The results achieved by semantic data analysis – a set of one-dimensional (string, sequential) patterns, or multi-dimensional patterns – is subject domain knowledge which is concentrated in a set of data signatures.

Further exploitation of the knowledge could include:

- classification – to identify in which known situation the process is operating,
- cluster verification – to check whether an existing classification is still valid,
- forecasting – to obtain a process trend for the future.

When considering water supply, data-guided knowledge is expected to be used in order to provide standard control and training scenarios. The cluster analysis approach was mainly used to produce a set of typical patterns – clusters – which represents variants in water consumption behavior. As previously mentioned, the water consumption profile analysis involves the sleep-wake cycle as an indivisible piece of information. The profiles available over a period of several months were assumed to possess information about the customer's – the water user's – behavior.

The clustering approach consists of two tasks: *first*, to structure the raw data into clusters; *second*, to map the clusters onto an a priori defined set of water user behavior scenarios. If the mapping is bijective, the clustering is successful. If the separability of the classes is high (which is given by the membership function), the used set of data profiles is informative and adequate for modeling the chosen set of water user behavior scenarios.

In Figure 4, the results of clustering with respect to the *a priori* expected behavior scenarios "workday" and "weekend" are depicted. The classification of the profiles was obtained by the *c-means* algorithm (the MATLAB software with Fuzzy Logic Toolbox (The MathWorks, Inc.)). The cluster set (Figure 4a) elaborated from a given set of data profiles will denote the signature of this set of profiles. In Figure 4b, the "workday" and "weekend" cluster centers as well as water consumption profiles over 2 months are shown as gray code matrices.

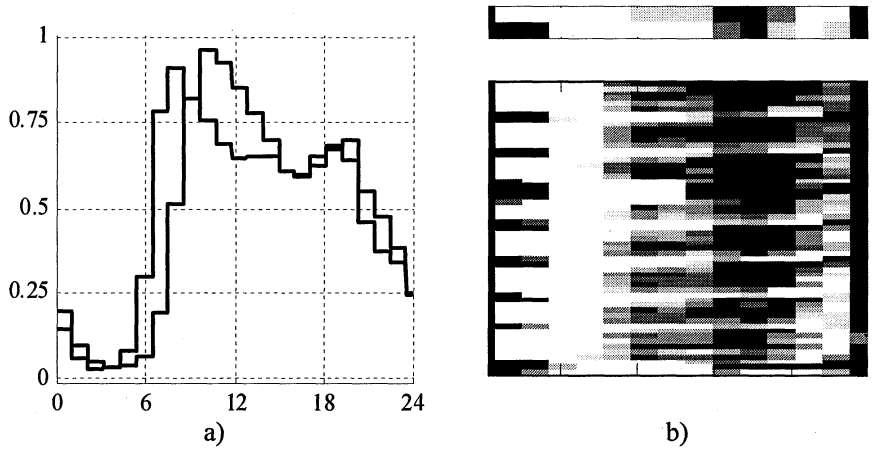

Figure 4 – Cluster centers for water consumption profiles over time axis:
a) "Workday" (solid line) and "weekend" (dashed line);
b) Above: "workday" and "weekend" cluster centers as gray code lines; Below: gray code lines for sleep-wake cycles over 2 months.

Here, matrix lines represent partial data profiles within the sleep-awake cycle. The superposed lines give impression about similarity of sleep-awake patterns over the workdays and weekends. In Figure 4b, the lines which correspond to weekends begin with a black zone which puts on view a delay in water consumption.

Considering the results of the performed data analysis, we recognize that the data signature has some relatively stable characteristic profiles for two typical situations: "workday" and "weekend". The operation scenarios of the technical facility can be elaborated for these two profiles respectively. The question of the initial operation scenario setup is solved using an a priori information – calendar. This represents the first exploitation scenario of the elaborated subject domain knowledge.

The deviation detection is based on analysis of distances between current process profile and cluster centers. Here, the membership function can be exploited. Moreover, also semantic aspects becomes an importance. By means of semantics reasoned weighting, a set of deviation functions can be provided. For calculus, different deviation functions may emphasize different properties of the data series.

Thus, the deviation identification for water consumption profiles has special semantic aspects which are listed below:

- profiles which are sequential patterns are being compared,
- time and duration of dynamic events (accelerations) are of importance;
- time and duration of maximal and minimal workload are of importance.

Moreover, the comparison of sequential patterns must have a highly robust results and be tolerant of the deviations in the range of intermediate workload.

Comparison of the codes proceeds using an algorithm for approximate string matching (Melichar, 1995). Obviously, an exact matching of two different patterns is impossible due to stochastic disturbances. Therefore, the procedure must reasonably allow, and also interpret, some "small" differences between the patterns which are being compared. The deviation calculus builds a measure for profiles similarity ("measure of confidence").

5. CONCLUSIONS AND OUTLOOK

The technique developed here was applied for analysis of water consumer behavior in German middle-sized cities. The examples shown here substantiate the approach for the modeling of the water consumers' behavior. The elaborated urban area specific set of signatures supports the facility operator as well as the manufacturer with new information which will be implemented in the control logic of the SCADA system. Moreover, this technique may build the functional interface for a data warehouse. In this context, the data warehouse model will provide the semantic background for the knowledge acquisition procedure (compare with (Kouba, 2002)).

The new application domain knowledge is already being used for the optimization of the facility operation. Thus, a simulation-based operator training with respect to typical operational situations is already available (Schreck, 2002). A model core for the water distribution facilities has already been implemented at the Fraunhofer IPK and can be supplied by data derived from the signature.

Further research activities are devoted to methods for data model adaptation at runtime, when new data samples are being imported. Herewith, a high sophisticated diversification of process control use cases should be achieved.

6. ACKNOWLEDGEMENTS

The study represented here was performed in the framework of an R&D project "Akquisition, Management und Integration von Prozesswissen in eine modellbasierte Prozessführung als Repräsentant einer neuen Generation von wissensbasierten Systemen in der Leittechnik" (http://amaryl.ipk.fraunhofer.de) partially funded by the Federal Ministry of Education and Research, Germany. We thank our industrial partner Elpro Prozessindustrie und Energieanlagen GmbH, Berlin, for extensive contacts with companies of the Water industry.

7. REFERENCES

1. Lisounkin A. "Semantic characterization of Data Series with Application to facility Control". In Proceedings of the IASTED International Conference on Signal Processing, Pattern Recognition, and Applications, June 30 – July 2, 2003, Rhodes, Greece, pp. 113-118.
2. Müller J.-A., Lemke F. Self Organizing Data Mining. Extracting Knowledge from Data. Libri Books on Demand, Dresden, Berlin, 2000.
3. Melichar B. "Approximate String Matching by Finite Automata." In Proceedings of the 6th International Conference on Computer Analysis of Images and Patterns (CAIP'95), Prague, Czech Republic, September 6-8, 1995.
4. Kouba Z., Matousek K., Miksovsky P.: "On-line analysis of utility networks", In Knowledge and Technology Integration in Production and Services: Balancing Knowledge and Technology in Product and Service Life Cycle. Edited by Vladimir Marik, Luis M. Camarinha-Matos and Hamideh Afsarmanesh, Kluwer Academic Publisher, 2002, p. 469-476.
5. Schreck G. "Simulation Services for Training of Plant Operators". In Knowledge and Technology Integration in Production and Services: Balancing Knowledge and Technology in Product and Service Life Cycle. Edited by Vladimir Marik, Luis M. Camarinha-Matos and Hamideh Afsarmanesh, Kluwer Academic Publisher, 2002, p. 79 – 86.