HUNGARIAN JOURNAL
OF INDUSTRIAL CHEMISTRY
VESZPRÉM
Vol. 35. pp. 75-84 (2007)

WHAT HAPPENS TO PROCESS DATA IN CHEMICAL INDUSTRY? FROM SOURCE TO APPLICATIONS – AN OVERVIEW

B. BALASKO, J. ABONYI

University of Pannonia, Dept. of Process Engineering, H-8201, Veszprem, P.O.Box 158, HUNGARY

It is globally accepted that information is a very powerful asset that can provide significant benefits and a competitive advantage to any organization, like production technologies in the chemical industry, which was driven by market forces, customer needs and perceptions, resulting in more and more complex multi-product manufacturing technologies. These technologies, due to their highly automated level, provide mountains of process data, which is applied only in daily operation and control, but it definitely can give access to the underlying structure of any system. To enhance this automation level while keep operation safe and efficient, one needs more information, i.e. knowledge about the process, which can be extracted from process data, and more tools, which can extract effectively this knowledge. To meet the growing expectations for future chemical engineering tasks, like multi-scale modelling, simulation and control or process and product design, advanced data analysis techniques can lead a way to solution. This paper briefly overviews some of the commercial products on market and the applicable data analysis techniques which guide process data from source to its application: from technology to expert knowledge with the help of knowledge discovery in databases (KDD) process. Numerous citations and their evaluation are given to show that data mining in chemical engineering can efficiently solve many data analysis related problems.

Keywords: process data, data analysis, data mining, chemical engineering, review

Introduction

Chemical engineering is said to be a profession of applied natural science, but besides applying the common practice for design, maintenance and control of industrial processes, it always faces challenges to continuously improve these techniques, thus improve the efficiency, effectiveness and reliability of all the chemical engineering activities. Charpentier defines the future main objectives of chemical engineering in four areas: (1) total multi-scale control of the process to increase selectivity and productivity; (2) equipment design of based on scientific principles and new operation modes and methods of production: process intensification; (3) product design; (4) an implementation of the multi-scale and multidisciplinary computational chemical engineering modelling and simulation to real-life situations [1]. It is clear that in every area, process data plays an essential role to fulfil these high expectations, hence it needs to be well structured and reliable.

The boom of the information systems in the past decades had its effect in every field of life, which is especially true for chemical industry, where a high level of automation and integration takes place. The high automation level provides the opportunity to collect more information (more variables) from the process and due to the integration of these components of the technology, the collected information in chemical industry can be larger than ever before. Additionally, due to the large developments in data storage capacities, the sampling frequency of the collected data has increased significantly as well. On the other hand, the availability of these modern data acquisition systems has increased as well: compared to a system 20-25 years ago, modern data acquisition systems cost 20 times less while running on higher performance level [2].

To serve this horizontal and vertical increase in data amount – it doubles every year – an exceptional hardware and software development takes place for a huge amount of application fields, and from being under-informed in the past we turned into over-informed: information mountains have arisen, but only ten percent of the enormous amount of collected and stored data is analyzed for further aims [3]. This means that there is a clear need for tools and applications that are able to handle all the relevant tasks regarding data produced in a process.

This paper aims to review the available solutions in the areas of data acquisition and data analysis for the above mentioned problems, highlighting the importance of process data analysis in chemical engineering. The first half of the paper presents recent solutions to data acquisition in industrial environments while the second half provides the various ways how these data can be analyzed to achieve process-related knowledge and meet the continuous development requirements in chemical industry.

Data acquisition and retrieval

The two main weaknesses of data acquisition systems are not handling heterogeneity and data inaccessibility:

1. Data from different sources and in different format cannot be handled in one environment, e.g. a priori knowledge, empirical or phenomenological knowledge cannot be incorporated into sampled data. Lots of research has been done on these problems: data compression and data integrity, the next section deals with several solutions to these problems.
2. A mid-size chemical plant has about few thousand measured variables sampled from seconds to hours, a hundred manipulated variables to control a few critical product quality related variables, which results in terabytes of data every year. It would mean inefficiently large data storage capacity if one wants to analyze not only prompt but historical data.

In this section solutions to these problems and already available commercial products on market are presented.

Integrated information storage and query

To solve the problem of heterogeneous data integrity several approaches have been developed. Complexity of integrating the information with their various describing models is not easy to handle, hence solution methods are different. Two main solution groups can be identified: where the integrality problem is solved at the query level or at the construction level of the integrated information system.

Collins et al. developed an XML based environment [4], while Wehr suggests an object-oriented global federated layer above information sources [5]. In [6], Bergamaschi et al. presents an object-oriented language as well with an underlying description logic, which was introduced for information extraction from both structured and semi-structured data sources based on tool-supported techniques. Paton et al. developed a framework for the comparison of systems, which can exploit knowledge based techniques to assist with information integration [7].

Another approach to handle the heterogeneity of information sources is the application of data warehouses (DWs) to construct an environment filled by consistent, pre-processed data [8]. The main advantage of a DW is that it can be easily adapted to a DCS and other process information sources of a process while it works independently. *Table 1* shows a comparison of a DCS related database and a data warehouse [9].

Table 1: Main differences of a DCS related database and a data warehouse [9]

	DCS related database	**Data warehouse**
Function	Day-to-day data storage for operation and control	Decision supporting
Data	Actual	Historical
Usage	Iterative	Ad-hoc
Unit of work	General transactions	Complex queries
User	Operator	Plant manager, engineer,
Design	Application-oriented	Subject-oriented
Accessed records	Decimal order	Million order
Size	100 MB-GB	100 GB-TB
Degree	Transactional time	Inquiry time
Region	Unit, product line	Product

Obviously, beside database integration among particular parts of the whole process, there is a need for information integration in the level of the whole enterprise as well for the purpose of optimal operation and planning. This task cannot be fully automated, there is a need for permanently improved methods and approaches for creation, storage and dissemination of experience, know-how and judgment embedded in the organization [10].

Appropriate time-series representation for data compression

Data compression is rather a contribution of the signal and image processing society where lossless information transmission is a key feature within limited time or bandwidth, in chemical engineering society data compression has beside storage capacity rationalization another important issue: retrieve the data in a manner that renders it easily interpretable for the execution of later engineering tasks. In this manner, data compression problem is turned into trend representation problem. Lin et al. gave a classification of process trend representation methods in [11], which can be seen in *Fig. 1*. Many of these representation techniques refer to segmentation of time series, which means finding time intervals where a trajectory of a state variable is homogeneous [12], representing data by its segments and storing only the segments instead of raw data.

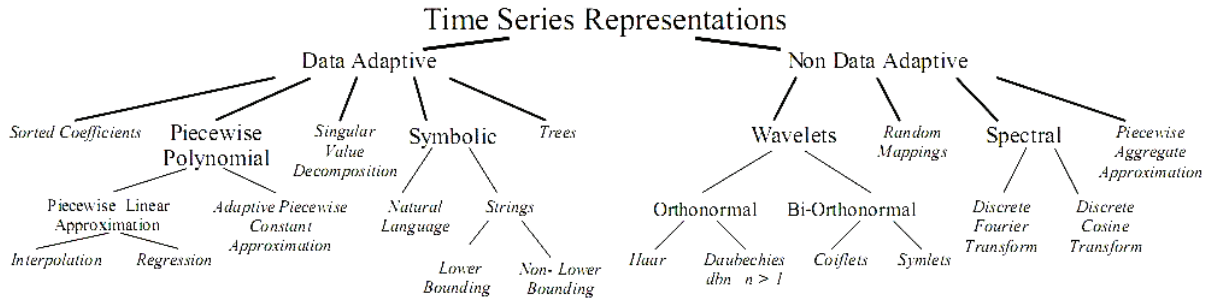

Figure 1: Hierarchy of various time series representations for data mining [11].

Products on the market

The modern distributed control systems (DCSs), which are widely implemented in modern, automated technologies have the direct access to the field instrument signals and measurements, while have data storage functions as well. Today several software products in the market provide the capability of integration of historical process data of DCS's: e.g. Intellution I-historian [13], Siemens SIMATIC [14], the PlantWeb system of Fisher-Rosemount [15], Wonderware Factory- Suite 2000 MMI software package [16] or the Uniformance PHD modul (Process History Database) from Honeywell [17], which structural components are shown in *Fig. 2*. These elements are typical in modern data collection systems.

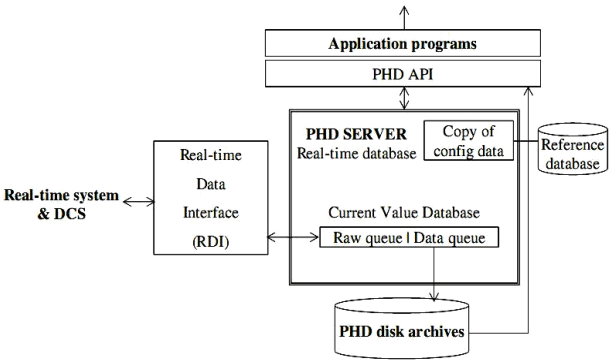

Figure 2: Structure of the data flow in Honeywell Uniformance PHD software.

There are two main operations:

- *Data collection*: Data originates from real-time system and is collected by a real-time data interface (RDI). Tag parameters for all the variables are stored in a reference database. A tag contains all important information about a process variable (name, type, unit, etc.). RDI sends data to PHD server which places the collected data for a tag in the raw data queue and applies data processing, such as smoothing, compression, and so on, to move raw data queue entries to the data queue of the tag. Data queue of the tag then holds processed data that is ready for insertion into the active logical archive files using the continuous store thread.
- *Data retrieval*: An application program makes a call to the PHD application programming interface (API) indicating the desired tag and time range for data. The PHD system checks the data queues to see if the data is still held in the queues, otherwise PHD accesses the data from the connected archive files.

Data flow goes as follows: First, the tag names of the relevant process variables are selected from all the possible tags in the plant. Process data belonging to the selected tags are accessed in PHD by the Uniformance Desktop application program (by Honeywell). While the Uniformance runs as an MS Excel add-in, the results of data queries are saved in Excel files.

Concluding, modern data acquisition systems need to be capable to handle diverse types of data in a way that data is applicable for further analysis. Rationally constructed data warehouses are needed for these purposes. Some of the above mentioned commercial historical data handling products assist DW maintenance interfaces as well, but in most cases there is no integrated software solution. Moreover, to get valuable knowledge that guides process development, appropriate information storage is not sufficient, process data analysis indispensable. The next section deals with this topic where a widely-applied procedure is presented.

Information extraction from process data

Knowledge Discovery in Databases (KDD)

Integration of heterogeneous data sources is highly related to knowledge discovery and data mining [18, 19], All in all this is one of its main purposes: store data in such a logically constructed way that some deeper information and knowledge can be extracted through data analysis. Knowledge discovery in databases (KDD) is a well known iterative process in the literature, which involves several steps that interactively take the user along the path from data source to knowledge [20].

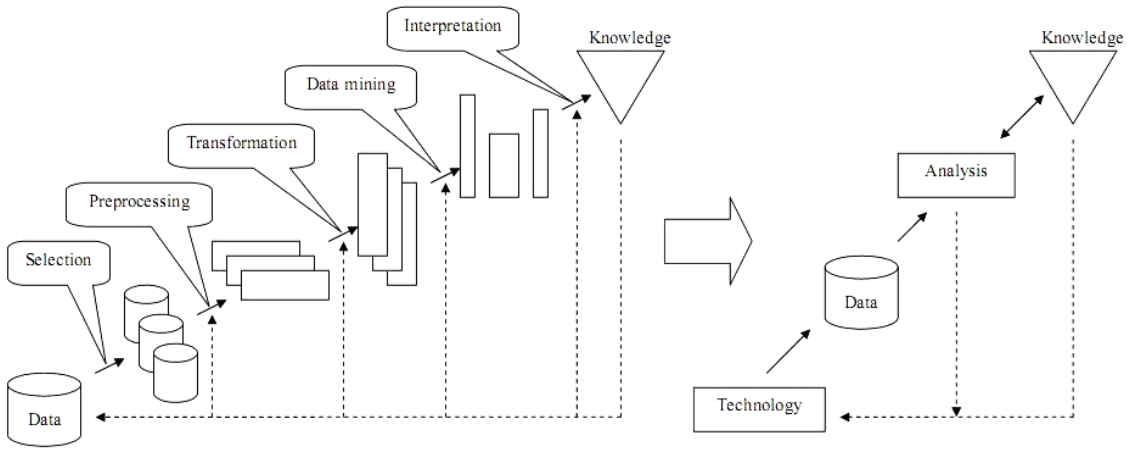

Figure 3: Knowledge Discovery in Databases process (left) and the data-driven process development scheme (right).

Fig. 3 shows the KDD process and its connection to the process development scheme: KDD can be considered as the analysis step of the process development process. This connection was published by many researchers who used the elements of KDD for solving several engineering tasks, like system identification, process monitoring and fault diagnosis, time-series analysis. In the following, we go through the steps of KDD highlighting the presence of *"data mining in chemical engineering"* (Note, that although data mining is a particular step of KDD, it is often associated to it as an independent technique).

1. ***Data selection.*** Developing and understanding of the application domain and the relevant prior knowledge, and identifying the goal of the KDD process.
2. ***Data pre-processing.*** This step deals with data filtering and data reconciliation. In process data warehouses and integrated KDD environments it is made preliminary during collection of relevant data.
3. ***Data transformation.*** Finding useful features to represent the data depending the goal of the task. Dimensionality reduction or transformation methods are applied to reduce the effective number of variables under consideration or to find invariant representation of data.

Data selection, pre-processing and transformation activities are often referred to as the ***data preparation*** step. It corresponds to the feature selection step of the pattern recognition process, which means to select a subset of original features that is good enough regarding its ability to describe the training data set and to predict for future cases. A wealth of approaches have been used to solve the feature selection problem, such as principal component analysis [21], Walsh analysis [22], neural networks [23], kernels [24], rough set theory [25, 26], neuro-fuzzy scheme [27], fuzzy clustering [28], self-organizing maps [29], hill climbing [30], branch and bound algorithms [31], and stochastic algorithms like simulated annealing and genetic algorithms (GAs) [32-33].

Process data have several undesirable attributes which need to be handled before any analysis can take place: time-dependent, multi-scale, noisy, variant and incomplete. All these problems need to be solved in the data preparation steps, hence it takes the largest part, approx. 60 % of the efforts in the whole KDD process. For industrial data reconciliation, OSIsoft and Invensys have developed packages such as Sigmafine and DATACON [34, 35].

4. ***Data mining.*** It is an information processing method, the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data (corresponds to feature extraction in pattern recognition).
   a) The goals of data mining are achieved by various methods:
   - *Clustering.* Cluster is a group of objects that are more similar to one another than to members of other clusters. The term "similarity" should be understood as mathematical similarity, measured in some well-defined sense. In metric spaces, similarity is often defined by means of a distance norm, which can be measured among the data vectors themselves, or from a data vector to some prototypical object of the cluster. The prototypes are usually not known beforehand, and are sought by the clustering algorithms simultaneously with the partitioning of the data. The prototypes may be vectors of the same dimension as the data objects, but they can also be defined as "higher-level" geometrical objects, such as linear or nonlinear subspaces or functions. Data objects belong to a cluster by their membership value, which is zero or one for hard clustering and between zero and one for fuzzy clustering techniques. Note, that in the case of fuzzy clustering the sum of the membership values equals one, i.e. a data object is more or less part of every cluster. On *Fig. 4*, clustering of data of a dynamic crystallizer cascade model (reconstructed in a 4-dimensional state space) projected by PCA is shown to analyze the cyclic operation [36].

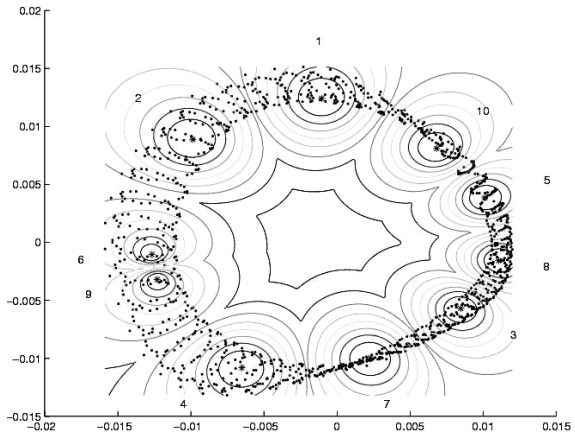

Figure 4: Fuzzy clustering of crystallizer cascade model data. Data points are denoted by dots, cluster prototypes by stars, cluster membership value levels by lines (darker means lower).

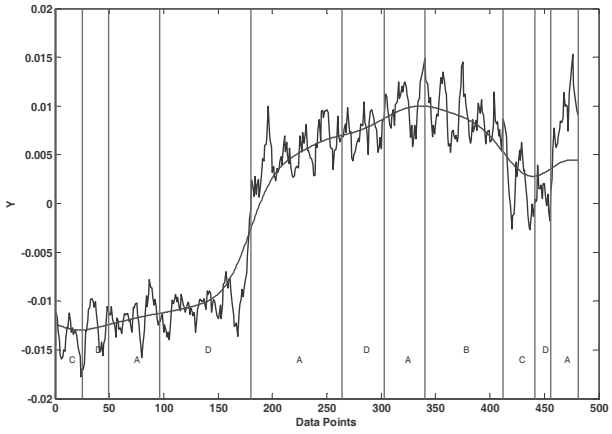

Figure 5: Second-order segmentation of filtered process transition data of polypropylene plant projected into 1-D by PCA. Segment boundaries are noted as vertical lines at extrema or inflexion points.

Clustering is widely used for feature selection [28], feature extraction method, which is applied in operating regime detection [36, 37], fault detection [38, 39] or system identification, like model order selection [40-42], state space reconstruction [43].

- *Segmentation.* Time series segmentation means finding time intervals where a trajectory of a state variable is homogeneous. In order to formalize this goal, a cost function with the internal homogeneity of individual segments is defined. This cost function can be any arbitrary function, usually it is defined by distances between the actual values of the time-series and the values given by a simple function (constant, linear or a polynomial function of a higher but limited degree) fitted to the data of each segment. Hence, the segmentation algorithms simultaneously determine the parameters of the describing models and the borders of the segments by minimizing the sum of the costs of the individual segments.

  The linear, steady-state or transient segments can be indicative for normal, transient or abnormal operation, hence segmentation based feature extraction is a widely known technique for fault diagnosis, anomaly detection and process monitoring or decision support [44-47].

  Fig. 5 shows a second-order segmentation of 1-D polymerization data during a process transition. Second-order means, segment borders are captured where the first or second derivative of a trend changes sign, thus at extrema and inflexion points.
- *Classification.* Map the data into labelled subsets, i.e. classes, which are characterized by their specific attribute called the class attribute. The goal is to induce a model that can be used to discriminate new data into classes according to class attributes. The induction is based on a labelled training set. The objective of the classification is to first analyze the training data and develop an accurate description or a model for each class using the attributes available in the data. Such class descriptions are then used to classify future independent test data or to develop a better description for each class. Many methods have been studied for classification, including decision tree induction, support vector machines, neural networks, and Bayesian networks [20]. In chemical engineering problems, classification is used in fault detection, anomaly detection problems [27, 45, 47-50]. On *Fig. 6*, a typical classification example is shown, where a decision tree was applied for the problem of the classification of operating regions related to the runaway of a chemical reactor. In [51] a new approach has been proposed, which is allows the transparent and interpretable representation of the boundaries of the operating regions.

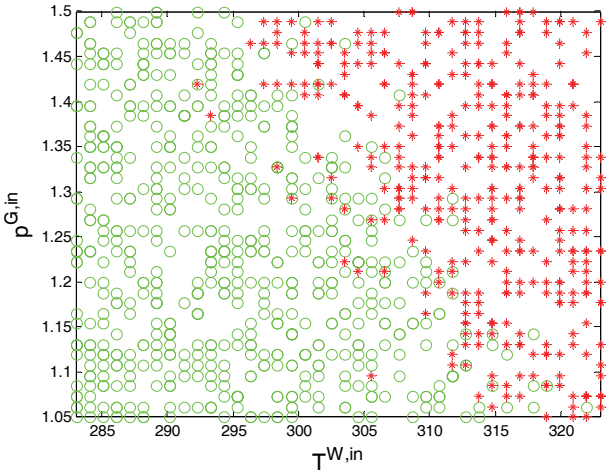

Figure 6: Classification example for the classification of operating conditions regarding the runnaway of a chemical reactor. The decision tree representation of the related classifier is shown in *Fig. 8*

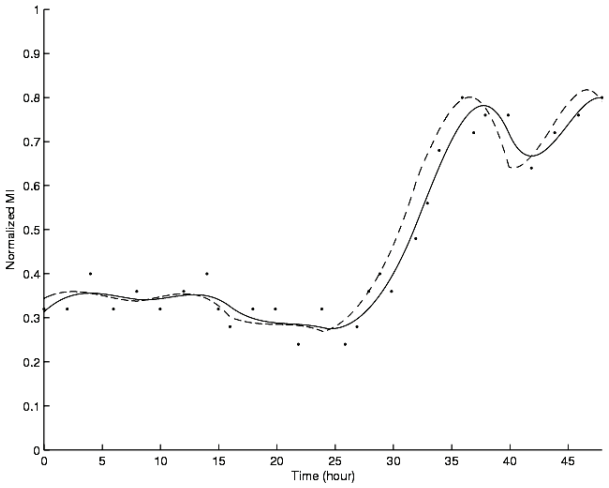

Figure 7: Cubic spline interpolation of a semi-mechanistic model for online Melt Index prediction in a polyethylene process.

- *Regression.* The purpose of regression problems is to give prediction for process or so called dependent variables based on the existing data (independent variable), in other words, regression learns a function which maps a data item to a real-valued prediction variable and the discovers functional relationships between variables [52-53]. Uses of regression include curve fitting, prediction (forecasting), modelling of causal relationships and testing scientific hypotheses about relationships between variables. Applied mainly in system identification problems, e.g. [54]. In [55], cubic spline interpolation-regression is applied to estimate variable derivatives for a semi mechanistic neural network model (*Fig. 7*).

b) Representation, i.e. output of data mining, of patterns of interest can be in form of several techniques as well:

- *Regression models.* Model interpretation of a system's behavior is possible by several techniques for numerous tasks. The extracted model structure can be various: from linear autoregressive models [53] to artificial neural networks [27, 45, 48], semi mechanistic models [37], self-organizing maps [29, 50], etc. On *Fig. 8* component planes of a SOM model for Melt Index prediction of a polypropylene polymer grade are presented for 8 independent variables of the technology [56].
- *Association rules.* General form of association rules is an 'IF X … THEN Y …' (noted as X□Y) implication. The two parts of a rule are the antecedent (X) and the consequent (Y). The association rules are constructed from frequent item sets [57]. The occurrences of an item (or item sets) in a data set are called support, which value could be seen as a probability value: how many percent of the transactions is the specific item (are the items of an item set together). An item is called frequent item if its support is higher than a given (user defined) threshold, namely the minimal support. The support of a rule is equal to the support of the item sets contained in the rule. While support says only the probability of joint occurrence of X and Y, the confidence (conditional probability) of an X□Y rule serves information about relationships between the X and Y.

  Association rules are applied in the field of decision support, process monitoring, process control [58].
- *Decision trees.* Common representation for classification problems [47, 49]. The goal of tree induction method is to get an input attribute partitioning which warrants the accurate separation of the samples. A decision tree has two types of nodes (internal and terminal) and branches between the nodes. The possible outputs for an internal node (cut) are represented by the branches. The terminal nodes of the tree are called leaves where the class labels are represented. The paths from the root to the leaves (sequences of decisions, or cuts) represent the classification rules. Therefore, as data partition representation, it represents the data as a hyper-rectangle. The most of the decision tree induction algorithms (e.g. ID3, C4.5) are based on the divide and conquer strategy. In every iteration steps the cut which serves topically the highest information gain (greedy algorithms) is realized.

  In *Fig. 9.* a decision tree is presented for reactor runaway detection of a fixed bed tube reactor [59]. There are two class attributes: class attribute 1 and 2 refers to reactor conditions where reactor runaway takes (1) and takes not place (2). Decision variables are: cooling water inlet temperature ($T^{W,in}$), reactor mixture inlet temperature ($T^{W,in}$), inlet pressure of reactor mixture ($p^{G,in}$) and mass feed flow of reactants ($B^{G,in}c_A{}^{G,in}$ and $B^{G,in}c_B{}^{G,in}$).

5. ***Interpretation of mined patterns*****,** i.e. discovered knowledge about the system or process. The interpretation depends on the chosen data mining representation.

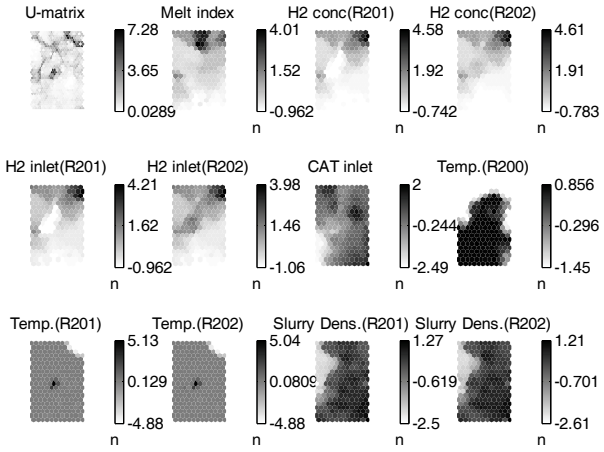

Figure 8: Self-organizing map representation based regression of process variables for Melt Index prediction in a polypropylene plant.

```
T^W,in <= 306.6 :
|  T^W,in <= 298.4 : 1 (331.0/10.8)
|  T^W,in > 298.4 :
|  |  p^G,in <= 1.37 :
|  |  |  B^G,in·c_A^G,in > 4.26 : 1 (86.0/5.1)
|  |  |  B^G,in·c_A^G,in <= 4.26 :
|  |  |  |  p^G,in <= 1.26 : 1 (14.0/5.5)
|  |  |  |  p^G,in > 1.26 :
|  |  |  |  |  B^G,in·c_B^G,in <= 3.02 : 2 (8.0/2.5)
|  |  |  |  |  B^G,in·c_B^G,in > 3.02 : 1 (3.0/1.9)
|  |  p^G,in > 1.37 :
|  |  |  T^G,in > 292.2 : 2 (28.0/4.8)
|  |  |  T^G,in <= 292.2 :
|  |  |  |  T^W,in <= 300.4 : 1 (3.0/1.9)
|  |  |  |  T^W,in > 300.4 :
|  |  |  |  |  B^G,in·c_B^G,in <= 1.26 : 1 (2.0/1.6)
|  |  |  |  |  B^G,in·c_B^G,in > 1.26 : 2 (4.0/2.1)
T^W,in > 306.6 :
|  p^G,in > 1.22 : 2 (165.0/9.4)
|  p^G,in <= 1.22 :
|  |  T^W,in <= 311.7 : 1 (30.0/8.6)
|  |  T^W,in > 311.7 :
|  |  |  p^G,in > 1.14 : 2 (34.0/2.9)
|  |  |  p^G,in <= 1.14 :
|  |  |  |  T^W,in > 317.9 : 2 (15.0/2.7)
|  |  |  |  T^W,in <= 317.9 :
|  |  |  |  |  B^G,in·c_A^G,in > 8.35 : 1 (6.0/2.4)
|  |  |  |  |  B^G,in·c_A^G,in <= 8.35 :
|  |  |  |  |  |  B^G,in·c_B^G,in <= 3.23 : 2 (17.0/5.7)
|  |  |  |  |  |  B^G,in·c_B^G,in > 3.23 : 1 (4.0/2.1)
```

Figure 9: Example of decision tree representation of a two-class problem for classification of reactor conditions regarding if there is (1) or is not reactor runaway (2). $T^{W,in}$: cooling water inlet temperature; $T^{G,in}$: reactor mixture inlet temperature; $p^{G,in}$: inlet pressure of reactor mixture; $B^{G,in}c_A{}^{G,in}$ and $B^{G,in}c_B{}^{G,in}$: mass feed flow of reactants.

For *visualization of the mined patterns*, Exploratory Data Analysis (EDA) has been developed. Although it is often stated as an independent analysis technique, it can be considered as a special application of the KDD process, where the knowledge is presented by the information embedded into several types of visualization tools. It focuses on a variety of mostly graphical techniques to maximize insight into a data set.

The seminal work in EDA is written by Tukey [60]. Over the years it has benefited from other noteworthy publications such as Data Analysis and Regression by Mosteller and Tukey [61], and the book of Velleman and Hoaglin [62].

Data preprocessing step in EDA refers to several projection methods in order to be able to visualize high dimensional data as well: techniques of principal component analysis (PCA) [63], Sammon-mapping [64], Projection to latent structure (PLS) [65], Multidimensional Scaling (MDS) [66] or Self-Organizing Map (SOM) [67] are applied. Data mining methods also use these techniques, but in EDA, projection is used for visualization purpose hence in most cases into two or three dimensions.

The graphical techniques of EDA have a wide spectrum including plots of raw data (histograms, probability plots, block plots), basic statistics (median, quantile plot, quantile-quantile plot, box plot) or advanced multidimensional plots (scatterplot matrices, radar plots, bubble charts, coded maps, etc.). In *Fig. 10*, *Fig. 11* and *Fig. 12*, some examples are presented.

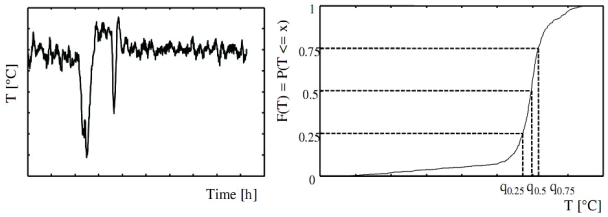

Figure 10: Example of a process variable (reactor temperature) and its cumulated distribution function ($q_{0.25}$, $q_{0.50}$, $q_{0.75}$ refer to quantiles) plotted by MATLAB

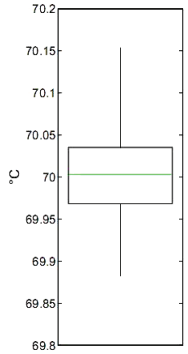

Figure 11: Box-plot of variable on Fig. 5 plotted by MATLAB, i.e. 5-number-summary from Tukey: minimum, maximum, median ($q_{0.50}$), 1st and 3rd quartile($q_{0.25}$ and $q_{0.75}$)

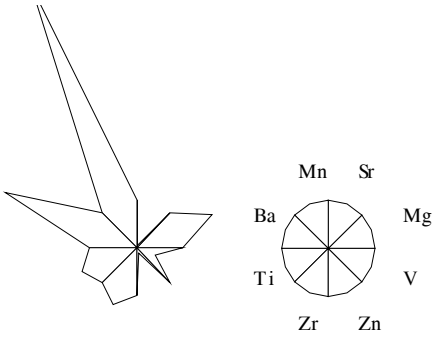

Figure 12: Star plot of a South African clinker (code number 159SA17). The standard on the right side can be used as comparison [68]

The most common software for EDA is MS Excel with free add-ins, but there are several products on the market as well: IBM's DB2 Intelligent Miner (which is no longer supported), Mathworks's MATLAB Statistics Toolbox [69] and the open-source WEKA developed by Waikato University [70].

Note, that most EDA techniques are only a guide to the expert to understand the underlying structure in the data in a visual form. Hence their main application is process monitoring [71, 72], but these tools are already used for system identification [73], ensuring consistent production [74] and product design as well [75].

Conclusions

Chemical industry is a highly automated industry, which produces a huge amount of production related data in every minute, which obviously has the potential to mine useful information and knowledge about the whole process. This paper reviewed how process data is stored and what types of scientific approaches are developed to guide this knowledge discovery.

The brief description of KDD and EDA techniques is presented, emphasizing their high correlation to chemical engineering tasks. From all the results in these scientific areas, one can conclude, that process data analysis has high contribution to the solution of problems that chemical engineers will face in the near future: optimal multi-scale control, process and product intensification, modeling and simulation of complex systems.

KDD gives users tools to shift through vast data stores to learn and recognize patterns, make classifications, verify hypotheses, and detect anomalies. These findings can highlight previously undetected correlations, influence strategic decision-making, and identify new hypotheses that warrant further investigation.

As it can be seen from the numerous citations, solutions based on the KDD process were proven to be extremely useful in solving chemical engineering tasks as well and showed that instead of simple queries of data, potential profit – through knowledge – can be mined by data analysis. The mined and discovered knowledge about the system or process is fed back to the beginning of the process to help continuous development (see Fig. 3).

REFERENCES

1. CHARPENTIER J. C.: Four main objectives for the future of chemical and process engineering mainly concerned by the science and technologies of new materials production, Chemical Engineering Journal 107, 3-17, 2005
2. AUSTERLITZ H.: Data acquisition techniques using PCs, Second edition, 2003
3. FAYYAD U., SIMOUDIS E.: Data mining and knowledge discovery. Tutorial Notes at PADD '97 – 1st Int. Conf. Prac. App. KDD & Data Mining, London.
4. COLLINS S. R., NAVATHE S., MARK L.: XML shema mappings for heterogeneous database access. Information and Software Technology, 44, 251-257, 2002.
5. WEHR H.: Integrating heterogeneous data sources into federated information systems. Proceedings of the 4th European GCSE Young Researchers Workshop, pages 1-11, October 2002. IESE-Report No. 053.02/E by Fraunhofer IESE.
6. BERGAMASCHI S., CASTANO S., VINCINI M., BENEVENTANO D.: Semantic integration of heterogeneous information sources. Data and Knowledge Engineering, 36, 215-249, 2001.
7. PATON N. W., GOBLE C. A., BECHHOFER S.: Knowledge based information integration systems. Information and Software Technology, 42, 299-312, 2000.
8. INMON W. H.: Building the Data Warehouse. John Wiley and Sons Inc., 3rd edition, 2002.
9. PACH F. P., BALASKO B., NEMETH S., ARVA P., ABONYI J.: Black-Box and First-Principle Model Based Optimization of Operating Technologies. In Proceedings of 5th MATHMOD, Vienna, 2006.
10. ZAHAYA D. GRIFFIN A., FREDERICKS E.: Sources, uses, and forms of data in the new product development process. Industrial Marketing Management, 33, 657-666, 2004.
11. LIN J., KEOGH E., LONARDI S., CHIU B.: A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. 2003.
12. KEOGH E., CHU S., HART D., PAZZANI M.: An Online Algorithm for Segmenting Time Series, IEEE International Conference on Data Mining, 2001.
13. CAPOCACCIA G.: Intellution production is the heart of manufacturing ebusiness, I-historian. Distributed Control Systems 7th Meeting, Miskolc, Hungary, 2001.
14. SEIDL U., SIMATIC Pcs 7: Efficient integration for tomorrow's DCS applications. Distributed Control Systems 5th Meeting, Miskolc, Hungary, 1999.
15. FÜLE S.: Integration of distributed - and enterprise control systems. Distributed Control Systems 5th Meeting, Miskolc, Hungary, 1999.
16. AJTONYI I., BALLAGI, A.: Integration of DCS in the complex producing system with wonderware factorysuite 2000 mmi software package. Distributed Control Systems 7th Meeting, Miskolc, Hungary, 2001.
17. GRINER, S.: New Rules of data management, InTech Magazin, Februar 2004.
18. GIANNADAKIS N., ROWE A., GHANEM M., GUO Y.-K.: Infogrid: providing information integration for knowledge discovery. Information Sciences, 155, 199-226, 2003.
19. SCOTNEY B., MCCLEAN S.: Efficient knowledge discovery through the integration of heterogeneous data. Information and Software Technology, 41, 569-578, 1999.
20. FAYYAD U., PIATESTKU-SHAPIO G., SMYTH P.: Knowledge discovery and data mining: Towards a unifying framework, Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1994.
21. MALHI A., GAO R. X.: PCA-based feature selection scheme for machine defect classification. IEEE Transactions on Instrumentation and Measurement, 53(6), 1517-1525, 2004.

22. SALCEDO-SANZ S., CAMPS-VALLS G., PEREZ-CRUZ F., SEPULVEDA-SANCHIS J., BOUSONO-CALZON C.: Enhancing genetic feature selection through restricted search and Walsh analysis. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, 34(4), 398-406, 2004.
23. VERIKAS A., BACAUSKIENE M.: Feature selection with neural networks. Pattern Recognition Letters, 23(11), 1323-1335, 2002.
24. SHIMA K., TODORIKI M., SUZUKI A.: SVM-based feature selection of latent semantic features. Pattern Recognition Letters, 25(9), 1051-1057, 2004.
25. JENSEN R., SHEN Q.: Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches. IEEE Transactions on Knowledge and Data Engineering, 16(12), 1457-1471, 2004.
26. SWINIARSKI R. W., SKOWRON A.: Rough set methods in feature selection and recognition. Pattern Recognition Letters, 24(6), 833-849, 2003.
27. CHAKRABORTY D., PAL N. R.: A neuro-fuzzy scheme for simultaneous feature selection and fuzzy rule-based classification. IEEE Transactions on Neural Networks , 15(1), 110-123, 2004.
28. MARCELLON F.: Feature selection based on a modified fuzzy C-means algorithm with supervision. Information Sciences, 151(5), 201-226, 2003.
29. YE H. L., LIU H. C.: A SOM-based method for feature selection. Proceedings of the 9th International Conference on Neural Information Processing, IEEE, 1295−1299, 2002.
30. FARMER M. E., BAPNA S., JAIN A. K.: Large scale feature selection using modified random mutation hill climbing. Proceedings of the 17th International Conference on Pattern Recognition. IEEE, 287-290, 2004.
31. SOMOL P., PUDIL P., KITTLER J.: Fast branch and bound algorithms for optimal feature selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(7), 900-912, 2004.
32. BHANU B., LIN Y.: Genetic algorithm based feature selection for target detection in SAR images. Image and Vision Computing, 21(7), 591-608, 2003.
33. OH I. S., LEE J. S., MOON B. R.: Hybrid genetic algorithms for feature selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(11), 1424-1437, 2004.
34. OSIsoft Inc., Sigmafine™, online access, URL: http://techsupport.osisoft.com/Products/Layered+Products/Sigmafine/Sigmafine+Overview.htm
35. Simsci Esscor, DATACON, online access, URL: http://www.simsci-esscor.com/us/eng/products/productlist/datacon/DATACON.htm
36. FEIL B., BALASKO B., ABONYI J.: Visualization of fuzzy clusters by fuzzy Sammon mapping projection – application to the analysis of phase space trajectories. Soft Computing, 11, 478-488, 2007.
37. ABONYI J., NEMETH S., VINCZE CS., ARVA P.: Process analysis and product quality estimation by Self-Organizing Maps with an application to polyethylene production, Computers in Industry, 52(3), 221-234, 2003,
38. ZOGG D., SHAFAI E., GEERING H. P.: Fault diagnosis for heat pumps with parameter identification and clustering. Control Engineering Practice, 14, 1435-1444, 2006.
39. PICIARELLI C., FORESTI G. L.: On-line trajectory clustering for anomalous events detection. Pattern Recognition Letters, 27, 1835-1842, 2006.
40. GARCIA C., BERNI C., NERI DE OLIVEIRA C. E.: Hardware/firmware implementation of a soft sensor using an improved version of a fuzzy identification algorithm. ISA Transactions, 47, 157-170 , 2008.
41. KILIC K., UNCU O., BURHAN TÜRKSEN I.: Comparison of different strategies of utilizing fuzzy clustering in structure identification. Information Sciences, 177, 5153-5162, 2007.
42. AZEEM M. F., AHMAD N., HANMANDLU M.: Fuzzy modeling of fluidized catalytic cracking unit. Applied Soft Computing, 7, 298-324, 2007.
43. LAN L. W., SHEU J. B., HUANG Y. S.: Investigation of temporal freeway traffic patterns in reconstructed state spaces. Transportation Research, Part C, 16, 116-136, 2008.
44. VENKATASUBRAMANIAN V.: A Syntactic Pattern-recognition Approach for Process Monitoring and Fault Diagnosis. Engineering Applications of Artificial Intelligence, 8(1), 35-51, 1995.
45. WONG J. C., MCDONALD K. A., PALAZOGLU A.: Classification of process trends based on fuzzified symbolic representation and hidden Markov models. Journal of Process Control, 8(5-6), 395-408, 1998.
46. SUNDARRAMAN A., SRINIVASAN R.: Monitoring transitions in chemical plants using enhanced trend analysis. Computers and Chemical Engineering, 27, 1455-1472, 2003.
47. CHARBONNIER S., GARCIA-BELTAN C., CADET C., GENTIL S.: Trends extraction and analysis for complex system monitoring and decision support. Engineering Applications of Artificial Intelligence, 18, 21-36, 2005.
48. ZHOU Y., HAHN J., MANNAN M. S.: Fault detection and classification in chemical processes based on neural networks with feature extraction. ISA Transactions, 42, 651-664, 2003.
49. ZHOU Y., HAHN J., MANNAN M. S.: Process monitoring based on classification tree and discriminant analysis. Reliability Engineering and System Safety, 91, 546-555, 2006.
50. YAN X., CHEN D., CHEN Y., HU S.: SOM integrated with CCA for the feature map and classification of complex chemical patterns. Computers and Chemistry, 25, 597-605, 2001.
51. VARGA T., ABONYI J., SZEIFERT F.: Applying decision trees to investigate the operating regimes of a

production process, Acta Agraria Kaposváriensis, (in press), 2008

52. POLLOCK D. S. G.: Classical Regression Analysis. Handbook of Time Series Analysis, Signal Processing, and Dynamics, 201-225, 1999.
53. RUSLING J. F., KUMOSINSKI T. F.: Analyzing Data with Regression Analysis. Nonlinear Computer Modeling of Chemical and Biochemical Data, 7-31, 1996
54. DAYAL B. S., MACGREGOR J. F.: Multi-output process identification, Journal of Process Control, 7(4), 269-282, 1997.
55. FEIL B., ABONYI J., PACH P. F., NEMETH S., ARVA P., NEMETH G., NAGY G.: Semi-mechanistic Models for State-Estimation - Soft Sensor for Polymer Melt Index Prediction. Lecture Notes in Computer Science, 3070 (2004) 1111-1117.
56. BALASKO B., NEMETH S., NAGY G., ABONYI J.: Application of integrated process and control system model for simulation and improvement of an operating technology. In Proceedings of European Congress of Chemical Engineering (ECCE-6), Copenhagen, 2007.
57. AGRAWAL R., IMIELINSKI T., SWAMI A.: Mining association rules between sets of items in large databases, Proceedings of the ACM SIGMOD International Conference on Management of Data, 207-216, 2003.
58. LIN C. T., LEE C. S. G.: Neural-network-based fuzzy logic control and decision system, IEEE Transactions on Computers, 40(12), 1320-1336, 1991.
59. VARGA T., SZEIFERT F., RÉTI J., ABONYI J.: Analysis of the Runaway in an Industrial Heterocatalytic Reactor, Computer Aided Chemical Engineering, 24, 751-756, 2006.
60. TUKEY J.: Exploratory Data Analysis. Addison-Wesley, 1977.
61. MOSTELLER F., TUKEY J.: Data Analysis and Regression. Addison-Wesley, 1977.
62. VELLEMAN P., HOAGLIN D.: The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis. Duxbury, 1981.
63. SMITH L. I.: A tutorial on Principal Component Analysis. 2002.
64. SAMMON J. W.: A Non-Linear Mapping for Data Structure Analysis, IEEE Trans. on Computers, C-18(5), 1969.
65. ZHAO S. J., XU Y. M., ZHANG J.: A Novel Nonlinear Projection to Latent Structures Algorithm, Advances in Neural Networks - ISNN Chapter 11, Springer Berlin / Heidelberg, 2004.
66. COX M. F., COX M. A. A.: Multidimensional Scaling, Chapman and Hall, 2001.
67. VESANTO J., HIMBERG J., ALHONIEMI E., PARHANKANGAS J.: Self-organizing map in MATLAB: the SOM toolbox, Proceedings of the Matlab DSP Conference, Espoo, Finland, 35-40, 1999.
68. TAMAS F. D., ABONYI J.: Trace Elements in clinker I. – A graphical representation, Cement and Concrete Research, 32(8), 1319-1323, 2002
69. The Mathworks Inc., Statistics Toolbox™, URL: http://www.mathworks.com/products/statistics/
70. WITTEN I. H., FRANK E.: Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
71. WANG D., ROMAGNOLI J. A.: Robust multi-scale principal components analysis with applications to process monitoring. Journal of Process Control, 15(8), 869-882, 2005.
72. URAIKUL V., CHAN C. W., TONTIWACH-WUTHIKUL P.: Artificial intelligence for monitoring and supervisory control of process systems. Engineering Applications of Artificial Intelligence, 20, 115-131, 2007.
73. MACGREGOR J. F., KOURTI T.: Statistical process control of multivariate processes. Control Eng. Practice, 3(3), 403-414, 1995.
74. MARTIN E. B., MORRIS A. J., PAPAZOGLOU M. C., KIPARISSIDES C.: Batch process monitoring for consistent production. Computers and Chemical Engineering, 20, 599-605, 1996.
75. LAKSHMINARAYANAN S., FUJII H., GROSMAN B., DASSAU E., LEWIN D. R.: New product design via analysis of historical databases. Computers and Chemical Engineering, 24:671-676, 2000.