appears in: *Proceedings of the International Symposium on Intelligent Data Analysis (IDA-95), Baden-Baden, Germany, August 17th-19th, 1995*

FACT-FINDING COMMITTEE WORK: DATA ANALYSIS BEYOND SINGLE SOURCES

Silvia Miksch
Austrian Research Institute for
Artificial Intelligence (ÖFAI)
Schottengasse 3, A-1010 Vienna, Austria

Johannes Gärtner
Abteilung für CSCW
Institut für Gestaltungs- & Wirkungsforschung
Technical University, A-1040 Vienna, Austria

ABSTRACT

Data analysis in real-world environments is not well defined. Data are usually more faulty than expected and the knowledge available is fuzzy and incomplete. Additionally, data analysis has to deal with different observation frequencies, different regularities, and different data types.
We propose a fact-finding process that takes the characteristics of real-world problems into account. All kinds of information available should be used to interpret the context. Therefore continuously and discontinuously numerical as well as qualitative data are used. Our approach consists of three main components based on temporal ontologies: data validation, data interpretation, and task adequate visualization. The data validation process classifies the input data according to their reliability. The data interpretation process leads to unified qualitative descriptions of point and interval data. Finally, these gathered and derived information is visualized task-oriented with more detailed descriptions on users' request to aid further refinements regarding validation and interpretation.
Our cyclic fact-finding process is applied in two different domains, namely artificial ventilation of newborn infants and shift-scheduling. The most crucial precondition of our approach is the task oriented structuring and interweaving of these three steps.

KEYWORDS

Data Validation, Data Abstraction, Visualization, Shift-Scheduling, Artificial Ventilation

INTRODUCTION

Theories on data analysis (e.g., Goebl and Schader 1992; Kay 1993) mostly work with well-defined problems. If one dares to work with real-world problems' one faces a bunch of data analysis problems. Different observation frequencies (e.g., high and/or low frequency data), different regularities (e.g., continuously and/or discontinuously assessed data), and different data types (e.g., quantitative and/or qualitative data) are somehow involved. Moreover, underlying structure-function models are in many cases poorly understood or not applicable because of their complexity and the fuzzy qualitative data involved (e.g., social or political constraints, qualitative expected trend descriptions). Therefore statistical analysis, control theory, or other techniques are often impossible, inappropriate or at least only partially applicable. However, as real-world-problems are real problems they have to be dealt with.
Therefore it is a central aim to present a task adequate global, comprehensive picture of all information available with respect to user demands, their experiences as well as to the degree of severity of a situation. Such a comprehensive picture may be achieved by context-sensitive interweaving of different knowledge-based approaches to classify input data and by adequate visualization techniques. More attention should be spend to this interweaving, even though it is a very complex and a partially domain specific task.

The guiding principle of our approach is to support a fact-finding process. All information available should be used to structure and to interpret the context. Finally, the gathered and derived information should be displayed task-oriented with more detailed descriptions on users' request. This approach is intrinsically cyclic. Its three main elements are: data validation, data interpretation (abstraction), and task adequate visualization. We applied our approach in two different domains: Firstly, monitoring and therapy planning of artificial ventilated newborn infants (VIE-VENT), and secondly, shift-scheduling (SPA).

VIE-VENT is especially designed for practical use under real-time constraints to support physicians in monitoring and therapy planning of artificial ventilation of newborn infants. According to the technical equipment of modern Intensive Care Units (ICU) a huge amount of on-line data is available. Additionally, off-line data and qualitative observations are available and needed for a global picture of the patient's condition and for an effective reasoning process.
SPA is designed to support shift-schedulers in developing long-term schedules for the duties of employees. These schedules have high impacts on working conditions, organization of work and costs. High numbers of partially fuzzy and (at the beginning of the scheduling process) ill-known data and requirements given by law, management, employee preferences, etc. have to be considered. The solution space is huge and from the scratch it is impossible to decide which requirements can be satisfied and which not (Gärtner 1992). Within the design process new questions regarding requirement refinement or relaxation arise and have to be solved. SPA is successfully in practical use in several plants.

In the following we will describe the three basic elements of our approach and illustrate the interweaving of these elements and their applicability using the two examples. Finally, we will discuss other areas where our framework can be applied.

DATA VALIDATION

The central aim of data validation is to detect faulty or contradictory input data and to arrive at classified data (e.g., reliable, inconsistent, unknown) for further analysis tasks (Miksch, et al. 1994). We apply a context-sensitive examination of the plausibility of input data based on different temporal ontologies to perform this analysis. Our data validation process uses discontinuously and continuously assessed numerical and qualitative data as well as derived qualitative descriptions. The latter are received from the data abstraction process described in the next section or are given by the users (e.g., user's requests). Applying these qualitative descriptions may result in a revision of the previous classified data that expresses one cyclic structure of our approach. We distinguish four data validation concepts based on their underlying temporal ontologies: time-point-, time-interval-, trend-based, and time-independent validation.

The time-point-based concept uses the value of a variable at a particular time point for the reasoning process. This concept can handle all kinds of data. It benefits from the transparent and fast reasoning process but suffers from neglecting any information about the history of the observed parameters. We apply range checking as well as causal and functional dependencies.
The time-interval-based concept deals with the values of different variables within an interval. We use three methods: temporal validity, allowed changes/values of a single variable during an interval and allowed changes/values of interdependent variables during an interval.
The trend-based concept tries to analyze the development of a variable during an interval. A trend is a significant pattern in a sequence of time-ordered data. Therefore the following methods can only handle continuously observed variables. It benefits from the dynamically derived qualitative trend-categories (descriptions) which overcome the limitations of predefined static thresholds. We apply trend-based functional dependencies of different dependent variables and an assessment procedure of the development of a variable.
The last concept is based on time-independent priority lists of variables and constraints. The data validation process allows to identify less reliable variables or constraints in case of conflicts. The result is a reliability ranking. This method is triggered, e.g., if an ambiguous classification of values (e.g., "some are wrong") has been derived.

Example VIE-VENT: Range checking
We enhanced range checking by adding attributes to define the clinical context. There are look-up tables for each variable with a variable name, a list of attribute descriptors, an upper and a lower limit. For example, ($pCO2$, (arterial, IPPV), 10, 140), where "arterial" refers to the kind of blood gas analysis and IPPV to the mode of ventilation. When a new variable value is received, the system checks if this value is in or out of the range and a corresponding flag is set. E.g., if $10 \leq pCO2(arterial, IPPV) \leq 140$ then it is a plausible measurement.
Example SPA: Time-interval-based
Time-interval-based validation methods are needed to check the historical data of operating and working hours (e.g., operating hours per week with respect to specific qualifications of the

workforce and with respect to work-places) and to check diverse scenarios for the assumed future demand. Even though one may assume these data are given in any company, real-world experiences showed that their acquisition and validation are quite difficult.

Example VIE-VENT: Trend-based functional dependencies & priority lists of variables
The increase/decrease of a variable suggests an increase/decrease of another variable after a delay-time. If such an expectation is violated, one of these variables must be faulty (classified as "some are wrong"). The reliability ranking of the variables suggests the more reliable variable. E.g., if the minute ventilation (AMV) is increasing then $P_{tc}CO_2$ is expected to decrease after 10 minutes.

DATA INTERPRETATION (ABSTRACTION)

The data abstraction process should lead to unified qualitative descriptions of point and interval data as well as verbal problem descriptions. The advantage of qualitative values is their unified usability in the data validation and visualization module, no matter of which origin they are. Adaptation to specific situations can easily be done by specific transformation tables without changing the model of data interpretation. According to our temporal ontologies we define three types of qualitative abstraction: time-point-, time-interval- and trend-based abstraction.
The time-point-based abstraction transforms quantitative data points into qualitative values. It is usually performed by dividing the numerical range of a variable into regions of interest. Each region stands for a qualitative category. The time-interval-based abstraction classifies a property of a variable to a time interval. A specific case of the previous abstraction mechanism is the trend-based abstraction, which classifies the development of a variable during a predefined time interval (Shahar, 1994).

Example VIE-VENT: data-point and trend-based abstraction mechanism
Figure 1 illustrates the guiding principle by the measurement of $P_{tc}O_2$ (transcutaneous partial pressure of oxygen). VIE-VENT smoothes oscillating data near transformation thresholds and performs a context-sensitive adjustment of data in clinical severe situation.
Data-point-transformation schemata relate numerical ranges to seven qualitative categories of blood gas abnormalities (qualitative *data-point*-categories listed on the left side of Figure 1). E.g., the transformation of the $P_{tc}O_2$ value of 91 mmHg during IPPV results in a qualitative $P_{tc}O_2$ value of *g3* ("extremely above target range").
The transformation of trend data into qualitative values is based on the combination of qualitative data-point-categories and the qualitative descriptions of the expected behaviors of a variable. In Figure 1 the *expected qualitative trend descriptions* (e.g., "variable $P_{tc}O_2$ is moving one qualitative step towards the target range within 10 to 30 minutes") are shown in the upper right corner.
These *trend-curve-fitting schemata* transform the quantitative trend values into ten qualitative categories (written in bold, capital letters in Figure 1).

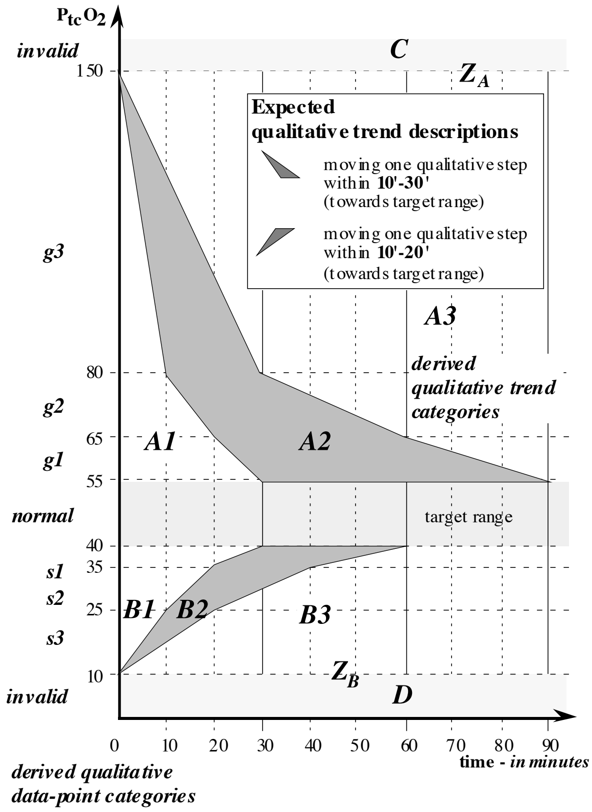

Figure 1: data-point- and trend-based abstraction of $P_{tc}O_2$

E.g., if a $P_{tc}O_2$ data point is classified as *g1, g2,* or *g3* (" ... above target range") we would expect a therapeutic intervention to result in a decrease of type A2 as "normal" trend (dark gray area in Figure 1). We apply a stepwise linearized algorithm to decrease the complexity of a comparison of exponential functions and to ensure responsiveness (Miksch, et al. 1995). These schemata are defined for all continuously assessed variables. These qualitative abstractions of continuously assessed variables make it easy to use simple rules to activate therapeutic actions.

Example SPA: time-interval abstraction

Time interval-abstraction methods are the core of the abstraction process. E.g., crucial constraints refer to employee related features of a schedule (e.g., the highest number of continuous resting hours within one week should be more than 36 hours). Therefore the combined effects of the starting-times of duties and their lengths with respect to resting hours are checked and the users are informed whether the schedule is o.k. Similar abstraction mechanisms are needed to check operating hours, weekly working hours, a high number of laws etc.

TASK ADEQUATE VISUALIZATION

The visualization module should display the derived qualitative and important quantitative values task-adequately and provide more information on users' request. Preparations of these data are necessary to display the case-relevant data in severe situations as comprehensive as possible (Arndt, et al. 1994; Gershon, 1994). The latter may result a re-processing of data validation and interpretation. This is the core of the cyclic and interactive structure of our approach.

Example: SPA

The high number of constraints to be considered does not allow to show all relevant information at the same time. The most important information for the design is displayed at workbench-level (see Figure 2). Further information can be shown easily using menus. The same approach is used for refinement and changes of variables and constraints, which may be done easily using specific buttons.

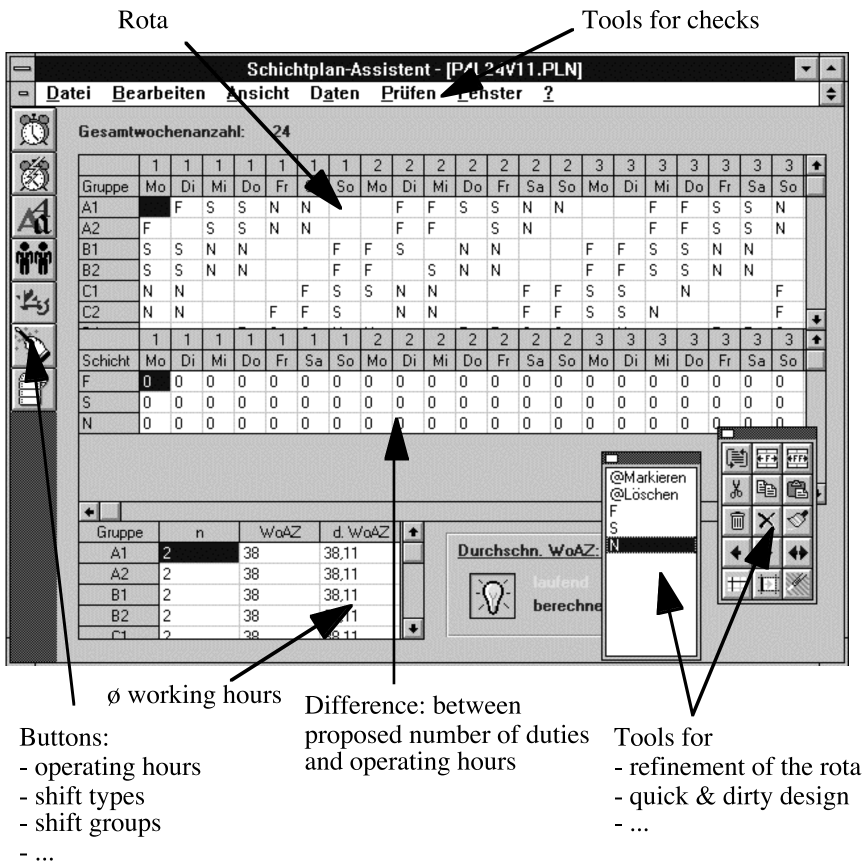

Figure 2: Screen shot of SPA

This work-bench supports cyclic refinements of the schedule and of the data/constraints considered.

The screen shot shows icons for refinements of basic features of the model at the left side (e.g., shift-definitions, operating hours, group definitions). In the middle there is a simple plan (i.e., group A1 starts with the morning shift, switches on Wednesday to an afternoon shift and on Friday to a night shift. The next two days are free). Differences to the operating hours needed (i.e., in this case no differences) and average working hours of the concerned groups of employees are shown. A toolbar and a menu for selecting shifts are shown on the right side.

Example: VIE-VENT

Observing multiple channels of data during longer time intervals results in extreme difficult tracking of all the data to realize significant changes or severe situations. Therefore these data cannot be handled within their raw structure. Our visualization module displays the derived qualitative values from the abstraction-process task-adequately and provides more information on users' request allowing re-processing of data validation and interpretation.

CONCLUSION: APPLICABILITY OF OUR APPROACH

Experiences with our two applications show that domain knowledge in the design of the system as well as in the actual situation of use is an essential precondition for effective data validation, interpretation and task-adequate visualization. Furthermore, it was crucial to support the fact-finding process, i.e. the cyclic and flexible refinement process based on the three elements of data validation, data interpretation and visualization. If they do work together well, as in the examples given above, data-analysis works even with real life problems.
The proposed framework of data validation, interpretation and visualization embedded into a cyclic fact-finding process driven by the needs of the user is applicable in time-related areas of planning and decision making with limited domain knowledge and real-world data (i.e., fuzzy, faulty, divers sources). The importance of the each method may vary depending on the problem characteristics as it did in the examples given above. An extension to other areas of reasoning may lead to an extension of the methods involved, still the basic approach of task-adequate interweaving should be useful.

REFERENCES

Arndt S., Lukoschek K., Schumann H. (1994); Design of a Visualization Support Tool for the Representation of Multi-dimensional Data Sets, in *5th Eurographics Workshop on Visualization in Scientific Computing*, Rostock, Germany

Gershon, N.(1994); From Perception to Visualization, in L. Rosenbaum et al (eds.); Scientific Visualization - Advances and Challenges, *Academic Press*, London (pp. 129-139)

Gärtner J.(1992); CATS-Computer Aided Time Scheduling - Ein Modell für die computerunterstützte (Schicht-) Arbeitszeitplanung, *Ph.D.Thesis*, TU-Wien

Goebl H., Schader M. (1992); Datenanalyse, Klassifikation und Informationsverarbeitung, *Physica*, Heidelberg

Kay S.M.(1993); Fundamentals of Statistical Signal Processing, *PTR Prentice Hall*, Englewood Cliffs, New Jersey.

Miksch S., Horn W., Popow C., Paky F. (1994); Context-Sensitive Data Validation and Data Abstraction for Knowledge-Based Monitoring, in Cohn A.G. (ed.), *Proceedings of the 11th European Conference on Artificial Intelligence (ECAI 94)*, Wiley, Chichester, UK (pp. 48-52)

Miksch S., Horn W., Popow C., Paky F. (1995); Therapy Planning Using Qualitative Trend Descriptions, in *Proceedings of 5th Conference on Artificial Intelligence in Medicine Europe (AIME 95)*, Springer

Shahar Y. (1994); A Knowledge-Based Method for Temporal Abstraction of Clinical Data, *Ph.D.Thesis*, Department of Computer Science, Stanford University

Acknowledgment

The current phase of the project VIE-VENT is supported by the "Jubiläumsfonds der Oesterreichischen Nationalbank", Vienna, Austria, project number 4666. We greatly appreciate the support given to the Austrian Research Institute of Artificial Intelligence (ÖFAI) by the Austrian Federal Ministry of Science and Research, Vienna.