MODERN TECHNOLOGIES OF IDENTIFICATION USING REAL TIME DATA WAREHOUSES [1]

Authors: ***S. HORAL, M. HORALOVÁ – KALINOVÁ, G. F. MICHAĽČONOK***
Workplace: ***Univerzita sv. Cyrila a Metoda v Trnave, Slovak republic***
Email: ***michaela.kalinova@ucm.sk, stanislav.horal@ucm.sk, german@ucm.sk***

Abstract:

Growing complexities of problems arising in different areas of real world lead to the necessity to apply a wide range of methods and tools for a deeper analysis of understanding the fundamental problem. Defining the fundamental problem helps to build such a model whose properties and behaviour would correspond to the nature of the problem. In the process of complex technical system modelling, great amounts of data of information value are supposed to be generated as a result of a wide range of experiments. Therefore, it is necessary to consider an optimal mode of data organisation. The aim of this contribution is to present ways of technical system identification using the MATLAB program, an interactive powerful program for scientific and technical calculations. We will also consider the possibilities of exploitation of data mining and data warehousing technologies in the process of identification.

Keywords: identification, data warehouse, real time

1. INTRODUCTION

The concept of system identification can be generally defined as a form of finding a suitable mathematical model of a corresponding system for a specific purpose. There exist two main views of the identification process, i.e., we come from the mathematical-physical analysis (analytical identification), or, we consider as a standard the measurements on systems (experimental identification). In this contribution we are going to present an experimental identification in a form of off-line identification, whereby we will assume that the measured values should be stored in suitable data stores represented by a data store and, subsequently, evaluated in the MATLAB environment. We will present the results of analysis of complex technological systems, i.e., complex systems whose input data and output data correspond to technological measurements.

The aim of this contribution is to present current automated facilities for identification of technological systems and to evaluate the possibilities of data mining and data warehousing technologies in this area.

2. REPRESENTATION OF EXPERIMENTAL DATA

We can define a complex system as a finite set of systems with mutual connections deployed in space. In identifying such complex system, particular attention is paid to the data flow analysis which distinctly defines connections between the elements of complex systems.

---

[1] This paper was supported by the VEGA Grant Agency, Grant No. 1/4078/07. The authors express their gratitude for the support.
Authors' addresses: Stanislav Horal, Department of Applied Informatics, University of SS Cyril and Methodius, Trnava, 917 01, Slovak Republic; Michaela Horalová-Kalinová, Department of Applied Informatics, University of SS Cyril and Methodius, Trnava, 917 01, Slovak Republic; German Michaľčonok, Department of Applied Informatics, University of SS Cyril and Methodius, Trnava, 917 01, Slovak Republic;.

With regard to dynamics of the complex systems, it is necessary to correctly determine the system parameters whose changes in time will be subject to research and which will represent the system state changes over time. The state of a complex system in time *t* can be defined as a set of immediate values of selected parameters, whereby we ignore the system's behaviour in time $t < t$. In terms of technological systems, the values of the parameters represent the values of quantities measured in individual places of an extensive system. On the basis of these assumptions, the dynamics of a complex system can be expressed as a sequence of system states in time. The immediate values of input quantities ***u(t)***, output quantities of the controlled system ***y(t)***, and the values of the measured state quantities ***x(t)*** represent the global state of the complex control system (Figure 1).

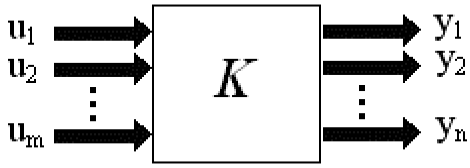

Figure 1 *A complex system*

3. DATA REPRESENTATION IN RELATIONAL DATABASE

In identifying difficult technological systems, we mostly come from the measured values of physical quantities. Such a wide data base needs to be represented by means of adequate data structures and saved in suitable data storages. The simplest way to save and process large volumes of data is using relational databases.

Relational databases represent a database system based on a relational model of data and relational algebra. The relational database model is formed by a set of normalised, time-variable database tables.

In general, the output signal of a dynamic system can be considered as a multiple (multi-channel) output signal. We can insert the values of an output signal into a separate database table so that we have available an *n-by-m* matrix, where *n* represents the number of values of output signal and *m* represents the number of output quantity (number of channels). The attributes of this database table may include, e.g., an identifier of a dynamic system, a timestamp or numerical marking of the order of dynamic-system state, sequential number of output quantity or its identifier, order number of value and the value of output signal of the dynamic system.

Analogically, the input signal of the dynamic system can be generally considered as a multiple input signal. Thus, we can apply the same manner to store the values of the incoming signal in the database table. The values of input quantities of the input signal and values of output quantities of the output signal can be stored in the form of real numbers with the required precision in accordance with application requirements, or in the form of whole numbers, tolerating a certain aberration. Basically, the information system does not have a data band width limit, but on the other side, we are interested in the speed of data processing. As the time of the data processing is minimal at the minimal band width of data, it is required to determine a minimal data band width from the aspect of metrological characteristics of data measurement. Metrological characteristics of data scanning can be the basis for establishing the information entropy and the dimension of information flow. In practice, there is no point storing the data with a larger data band width than we have determined from the information flow.

Another possibility is to hold serialised values of one input or output quantity in one particular state of the dynamic system in the database tables, which is not in conformity with normal forms, and even introduces additional costs of data processing, e.g. in the process of data extraction, transformation and loading into the date warehouse (ETL).

4. USING REAL-TIME DATA WAREHOUSES

A data warehouse is an organisation's repository of subject-oriented, integrated, time-variable and historical data used for gaining information and supporting decision making process. The data store contains atomic data and aggregate data.

The data present in a data store are integrated from several, typically non-homogenous sources, whereby the integration process includes data transformation into a uniform resultant format, data restructuring into adequately defined structures as well as data quality control. The data stored in a warehouse represent the states and values of real processes in a particular time instant.

When the data are loaded into the real-time data warehouse, the data are loaded in real time. Therefore, it is not possible to plan continual technical pauses for data loading and, at the same time, it is necessary to keep the data in a consistent state. In the recent time, with an increasing demand for solutions to some problems and analyses in real time, a series of advanced ETL tools ensuring performance of a particular ETL process in real time have emerged on the market. A solution, when the ETL process does not run in real time but only in almost-real time, is applicable to the dynamic systems in which the ETL process and a subsequent analysis are at least twice as fast as the dynamic system control frequency. If this requirement is not fulfilled for the whole range of frequency control, it is possible to use only lower frequencies for the dynamic system control and to use advanced frequencies for prediction.

The benefits of utilisation of real time data warehouses include not only the adapted data store for keeping great amounts of historical data, but also the availability of analytic services and tools provided by the data warehouses. The OLAP (Online Analytical Processing) tools perform analytical and reporting activities which can be utilised in the dynamic system identification.

Data mining is a process of data analysis from different aspects and data transforming into usable information. From mathematical and statistical aspects, the focus is on searching correlations, i.e., determining interrelations and patterns hidden within the data. In this way, we can reveal new dependencies between the measured quantities, which could be hardly discovered by classical identification methods.

Data mining parts of analytical services are being enhanced with additional algorithms for every new version. The SQL Server 2008 has 12 algorithms currently available and the Oracle Data Mining has also 12 data mining algorithms which are currently available.

Most suppliers of analytical tools include the tools with graphical interface for the visual design of a prediction query. With the help of the user's guide it is possible to select the data mining model, the algorithm used, the data source, the database columns on which the results of analysis depends, the database columns whose values we want to predict, and other information. A visual design results in a difficult and complex prediction query. Similar to the commands for creating OLAP cubes and filling in the tables of facts and dimensions with data, the prediction query is also described by the SQL language that can be used in any other database application. For our purposes, the results of prediction query can be processed using the MATLAB environment.

5. REPRESENTATION IN MULTIDIMENSIONAL STRUCTURE

For representation of experimental data by means of a multidimensional structure of the data store, it is optimal to represent the complex system in the state space. Our proposition comes from a discrete-state model in a standard form:

$$
\begin{aligned}
x(k+1) &= A.x(k) + B.u(k) \\
y(k) &= C.x(k) + D.u(k)
\end{aligned}
,
$$

where ***x(k)*** represents a column vector of state quantity values, ***u(k)*** is a column vector of input quantity values, and ***y(k)*** is a column vector of output quantity values in the k-th moment of measurement. For the detection of mutual internal connections and correlations amongst the quantities of a complex system, each relevant measurable variable must constitute a separate dimension of the multidimensional structure. The time dimension of a multidimensional structure is a bearer of values representing discrete-time measurements.

6. CONCLUSION

In this contribution we have tried to examine the utilisation of data warehouse and data mining technologies providing large possibilities of data collection and data representation and offering suitable techniques to acquire knowledge on data interactions that can be employed in the identification process. We have tried to propose an optimal multidimensional structure to represent the values of measured quantities in the complex dynamic system. We have also presented identification possibilities of the tool designed for scientific and research calculations and suggested the possibilities of co-operation between the MATLAB system and the database interface allowing access into the multidimensional structure of data warehouse.

Acknowledgement

This research has been supported by the VEGA Grant Agency, Grant No. 1/4078/07. The authors express their gratitude for the support.

References

[1] BRUCKNER, R. M., JENG, J. J., SCHIEFER, J. *Real-time Workflow Audit Data Integration into Data Warehouse System.* ECIS, Naples, 2003.

[2] ADELMAN, S., MOSS, L. T. *Data warehouse project management.* Addison-Wesley Longman Publishing Co., Inc. Boston, 2000.

[3] HUDZOVIC, P. *Identifikácia a modelovanie.* Bratislava, ES SVŠT, 1986.

[4] SÖDERSTRÖM, T., STOICA, P. *System identification.* Prentice Hall Int., 1989.