ISSN 1009－3095 *Journal of Zhejiang University SCIENCE* V.3, No.5, P.538－542, Nov.－Dec., 2002
http://www.periodicals.com.cn; http://www.zju.edu.cn/English
http://www.zjupress.com; http://lib.zju.edu.cn/eindex.htm; jzu_s@mail.hz.zj.cn

Data-mining massive real-time data in a power plant: challenges, problems and solutions*

CHEN Jian-hong(陈坚红)†, REN Hao-ren(任浩仁),
SHENG De-ren(盛德仁), LI Wei(李蔚)

(*Institute of Power Plant Thermal Energy Engineering & Automation, Zhejiang University, Hangzhou 310027, China*)

†E-mail: EnergyStar@cmee.zju.edu.cn

Received Oct.11, 2001; revision accepted Jan.21, 2002

Abstract: Nowadays, the scale of data normally stored in a database collected by Data Acquisition System (DAS) or Distributed Control System (DCS) in a power plant is becoming larger and larger. However there are abundant valuable knowledge hidden behind them. It will be beyond people's capacity to analyze and understand these data stored in such a scale database. Fortunately data-mining techniques are arising at the historic moment. In this paper, we explain the basic concept and general knowledge of data-mining; analyze the characteristics and research method of data-mining; give some typical applications of data-mining system based on power plant real-time database on intranet.

Key words: Data-mining, Power plant, Database, Real-time, Intranet
Document code: A **CLC number:** TK233, TP274

INTRODUCTION

Nowadays, the data normally stored in a database collected by Data Acquisition System (DAS) or Distributed Control System (DCS) in a power plant is usually gigabytes, even hundreds of gigabytes. For example, the real-time data collected in a medium-scaled power plant can be several gigabytes in a single month. Modern computer and database technology have the capacity for storing, processing, and rapidly searching for data in a database of this scale, and transform "data flood" into "ordered mass data collection".

However, most power plants use only traditional data analysis and simple traditional statistical methods to collate and analyze data, etc. Limitation of manpower, material, financial resources, and knowledge, preclude deep understanding and effective use of data in spatiotemporal sense. Thus the data cannot be optimally used. On the contrary they brought about a series of "data calamities" and "resource dilapidation". On one hand a power plant invests a lot of manpower, materials and financial resources, in constructing DAS, DCS and MIS but the efficiency of the traditional data analysis method is impaired significantly in view of the huge volume of data and as the inaccessibility of profound information hidden inside the data limits the improvement of management level and enhancement of economic profits. On the other hand, the power plant decision-makers hope to make correct decisions with the support of the information contained in these data but because of the knowledge limit, they cannot mine valuable information from this "data treasure house". It is just like the saying of John Naisbett that "human beings are submerged by data, but they are still longing for knowledge."

The power plants need for rapid data accessing and collecting techniques to replace the tradition crude data analysis methods moved us to find a technique, a new approach to intelligently collect and analyze data automatically so that the "data treasure house" can be optimally utilized. This is the data-mining, which will be dealt with in this paper.

* Project (No.06－020017) supported by Zhejiang Provincial Electric Power Corp.

CHARACTERISTICS OF DATA-MINING IN REAL-TIME DATABASE

The process of finding hidden information from data has different definitions including: knowledge extraction, information discovery, data-mining and the more proper knowledge discovering in database (KDD). In particular, data-mining can be considered a specific phase of the KDD process, which also includes other activities, such as data selection, checking, cleaning, preparation, pattern presentation and knowledge refinement and visualization (Fayyad et al., 1996a).

Data-mining is a process of extracting the pattern from the data. The knowledge coming from data mining is generally presented as a concept, rule, discipline. Usually data mining is regarded as an important step in the process of knowledge discovery in database. Data-mining and KDD are the outcome of unifying multi-disciplines, such as database technology, artificial intelligence, machine learning, statistic analysis, fuzzy logic, artificial neural networks and so on (Fayyad et al., 1996b).

Along with the electric power industry development, more and more large-scale production units have been put into operation in our country and all of these large-scale production units are equipped with the advanced DCS control system which can provide comprehensive functions of information collection, transmission, processing, storing, query and control, and can carry out online performance analysis for the units. Moreover, along with the going deep into information based construction, many electric power plants have established Intranet and implemented interlinkage with DCS. This allows storing of data in the Intranet's database server in the form of database for real-time data collection and disposal by DCS, which builds up a "data treasure house" in the process of production. But unfortunately at present, the application of "data treasure house" is confined within classing and analysis for traditional data and obtaining their surface information such as table statistic, trend analysis and so on while the internal relations of the data's attributes and implicit information are not accessible, viz., it cannot provide more important information for the ever changing market economy or continuous operational condition, and also cannot provide objective and predictively scientific basis for important decision-making and optimum operation of the power plant. Therefore, studying Data Mining & KDD based on power plant real-time database on Intranet, discovering the deep-seated information of intranet-based database, such as the rules of energy conversion, utilization and loss in the thermal equipment, the rules of gradual change of performance state of thermal equipment and shortened useful life. And these applications certainly will enhance the competitive ability and bring about considerable economic and social benefits for the enterprises in the present condition of the market economy.

The study of data-mining & KDD based on Intranet real-time database is faced with some problems as follows (Chen et al., 2001b):

1. Studying the causes of data loss or distortion error, selecting the proper ones from the miscellaneous data and wiping off noise and interference.

2. Studying the data-mining from real-time data because of their incessantly changing characteristics.

3. Studying database management system (DBMS) improvement and the optimized database design in favor of data-mining.

4. Studying the algorithm of data-mining & KDD.

5. Studying data-mining & KDD of Intranet-based real-time database (containing various data types such as integer, real, boolean and so on).

6. Studying the natural language or visualization technology in favor of comprehension and expression of discovered knowledge which would be advantageous to decision-making.

The approaches to studying data-ming & KDD of Intranet-based real-time database in power plant are as follows (Fayyad et al., 1996b):

Bayes decision theory This is a theoretical method for data-mining based on conditional probability, viz., selecting some subsets from attributes sets according database and finding the relation between conditional probability and attribute probability of decision-making attributes.

Artificial neural network This is a network

composed of artificial nerve cells imitating human brain structure and is also a nonlinear predictive and sorting model which had experienced training and learning. It can perform classing, clustering, characteristic mining, and so on.

Rough collection theory It is a mathematical tool that can process ambiguous and uncertain problem. The problems that can be processed with the use of this theory include data predigestion (removing redundant data), discovering the relativity of data, evaluating the meaning of data, deriving the algorithm of decision-making from data, approximate analysis of data, discovering approximation or difference of data, discovering normal forms and relation of cause and effect in data (Pawlak, 1998).

Fuzzy logic It combines the conceptions of fuzzy collection and Boolean logic. The truth value of a formulation can be either value in the interval of [0,1] which is always used in evidence synthesizing and confidence calculation in the study of data-mining & KDD.

In addition, the decision tree, the genetic algorithm and the visualization technology, and so on are all in favor of further research work.

SOME APPLICATIONS OF DATA-MINING

Some applications of data-mining & KDD based on power plant real-time database on Intranet are as follows:

1. Combined cycle power plant generator cost characteristics modeling

Combined cycle power plants are of great interest in many countries due to their high efficiencies and their low investment costs. In general, the working unit of these plants consists of two gas turbines, two boilers (heat recovery steam generator--HRSG) and a steam turbine to generate electricity. A typical combined cycle unit configuration is shown in Fig. 1.

Due to the highly non-linear behavior of gas turbine, HRSG and steam turbine, non-linear models must be used to represent the combined cycle power plant generator cost characteristics model. In this case, the method of data-mining based on combined cycle power plant real-time database on Intranet are used.

There are challenges in mining such a large time series database. Obvious ones are the size of the data which is always of concern, and the time dimension. Because of the high dimensionality and the varying phase factor of the time series, it is impractical to mine the data in the time domain. The classification has to be performed in some feature domains where their dimensions are low. The costly transformation of a large number of time series to the feature domain is inevitable. Selection of suitable features for appropriate representation of these time series in terms of conciseness and accuracy is also a research problem. It is evident that the data has problems. Noise, invalid measurements and missing values exist in every time series. To solve these data problems a big computational cost has to be paid to preprocessing large time series data. Data preprocessing is not only an extremely important but also very time-consuming stage in the knowledge discovery in databases (KDD) process. This is true in our analysis. The major data problems that need to be solved in data preprocessing are noise, invalid and missing values in time series. Noise can overwhelm signals and missing values can result in big gaps in time series. Invalid values can be easily identified and deleted from the sample data. In the following we present the data preprocessing techniques that have been used to handle noise and missing values (Chen et al., 2002).

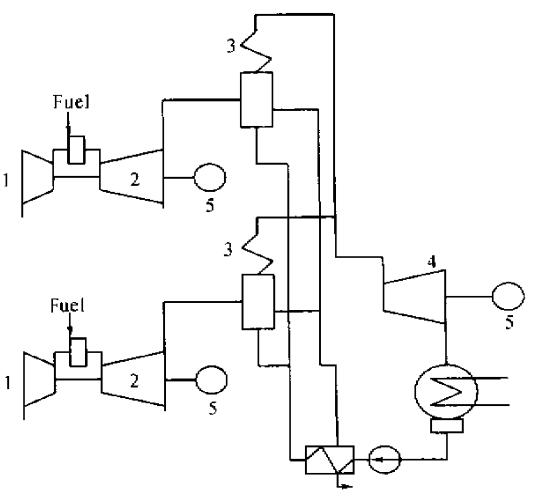

Fig. 1 The typical combined cycle power plant configuration

1 compressor; 2 gas turbine; 3 HRSG; 4 steam turbine; 5 generator

(1) Noise

Noise in a signal is composed of random noise and system noise. Noise levels are different in different signal according to their characteristics. The estimation of noise is represented as error-bar of E_i in the data set. If the noise of a signal is so large that it overwhelms the value of signals, the value of the signal will become uncertain and thus effectively unobservable.

A filter is designed to sieve out the true value from the real-time database. However, the values in which signals were overwhelmed in noise were removed from the sample data.

(2) Missing values

Missing values cause problems when they appear as big gaps in the time series. These big gaps add uncertainty to the estimation of values. To reduce the effect of big gaps we designed another filter to sieve out the true value from the real-time database (Chen et al., 2001a).

Table 1: shows the available operating data for a combined cycle power plant having two gas turbines (GTs) and one steam turbine (ST).

Table 1 Operating data of a combined cycle plant

Output power (%)	Unit status	Output power P (MW)	Heat rate F (GJ/h)
50	1GT(full)+ST	150.0	1297.65
75	2GT+ST	225.0	2017.72
100	2GT+ST	300.0	2532.00
	1GT(open cycle)	100.0	1255.45

After mining the real-time database on Intranet, we reached the following conclusion. Beyond 150MW, the second gas turbine must start on open cycle operation until it supplies enough exhaust heat to create good quality steam in the heat recovery steam generator (HRSG). When $207.353 \leqslant P < 225$MW, the power increase is due to increased output of the steam turbine only, while the gas turbine operates at a fixed output. For $P \geqslant 225$MW, all three turbines operate in closed cycle. Based on the above data and the operating conditions, the heat characteristics are derived as given by the following equation and shown in Fig. 2. Due to the availability of intermediate operating points, it was approximated by piecewise linear curves.

$$F=\begin{cases}12.555P-585.6; & 150\leqslant P<207.353\text{MW}\\ 2017.72; & 207.353\leqslant P<225\text{MW}\\ 6.857P+474.9; & 225\leqslant P<300\text{MW}\end{cases}$$

The above heat rate characteristics were multipled by the fuel cost to obtain the cost characteristics.

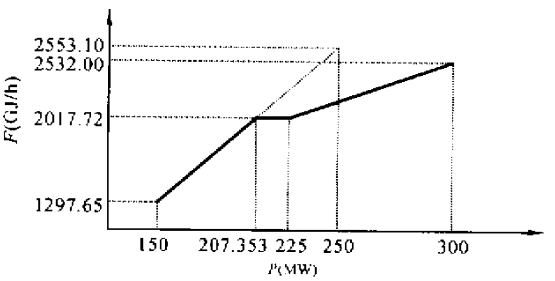

Fig. 2 The heat rate characteristics of a combined cycle power plant

2. Performance monitoring and state diagnosis

At present, the power enterprises in our country are in the period of managing pattern conversion from planned running to commercial running. There are some especially increasingly imminent requirements for production units' variable operation and state inspection, performance monitoring and state diagnosis. From studying data-mining & KDD based on power plant real-time database on Intranet, we can obtain valuable and optimal operation rules and running states information and complete equipment state-diagnosing appraisal and finally implement predictive maintenance which was advocated many years ago but never came into being until now (Chen et al., 2001b).

3. Energy-loss-diagnosis of thermal equipment

These days, the reform of power-price bid on network is implemented on all power networks which require high level of production units' economic operation and profits. Improved from the functions of original heat-engine units performance monitoring & analyzing system, with the study of data-mining & KDD based on power plant real-time database on Intranet, valuable and better energy conversion rules can be obtained, energy loss diagnosis of thermal equipment can be done and optimal operation and op-

eration direction can be implemented (Chen et al., 2001b).

SUMMARIZATION & DISCUSSION

Data-mining & KDD is a newly grown up research field and with the use of various an important topic requiring resolution pushed forward along with the development and prevalence of artificial intelligence, computer science & technology and web technology. The research of data-mining & KDD based on power plant real-time database put forward by the author can refine the abstract knowledge from a mass data collection by the use of various theories and approaches and thereby reveal the internal relations and essential rules hidden inside these data. As a new concept, data-mining & KDD:

1. It achieved a breakthrough in the application goal and the level of power plant management information system and is aimed at the practical appeal of enhancing production, and improving management. Being applied in power plants which convert the energy of chemical fuel to electric energy, the data-mining & KDD can reveal profound rules for energy conversion; improve the conversion efficiency and bring about better economic & social profits for enterprises under the market economy condition.

2. It is multi-disciplined because of its applying computer technology, artificial intelligence, artificial neural networks and heat-work conversion theory.

3. The research outcome is considerably practical and instructive.

References

Chen, J.H., Sheng, D.R., Li, W., Ren, H.R., 2001a. On-line forecasting-validating model of real-time data for turbogenerator operating expert system. *Power system engineering*, **17**(6):375－378.

Chen, J.H., Ren, H.R., Sheng, D.R., Li, W., 2001b. Investigation on knowledge discovery and data mining based on the real-time turbogenerator's monitoring data. *Zhejiang electric power*, (6):7－10.

Chen, J.H., Li, W., Sheng, D.R., Ren, H.R., 2002. A data fusion method for on-line performance calculation of turbogenerator. *Proceedings of the CSEE*, **22**(5):152－156.

Fayyad, U., Uthurusamy, R. 1996a. Data mining and Knowledge Discovery: Making Sense Out of Data. IEEE Expert, Oct, p.20－25.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996b. From Data Mining to Knowledge Discovery: An Overview. *In*: Faygad U, ed. Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, p.1－34.

Fayyad, U., 1996c. From Data Mining to Knowledge Discovery: Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press.

Guo, Y., Wang, Y., 1998. Data mining and knowledge discovery in database: a survey. *Patten recognition & Artificial Intelligence*, **11**(3): 292－299.

Pawlak, Z., 1998. Reasoning about data-A rough set prespective. LNAI 1424, Proceeding of RSCTC'98, Warsaw, Springer, **6**:25－34.

Wang, L.Q., Tang, C.J., Yu, Z.H., He, X.M., 1998. Web-based Data Mining. *Compumter applications* **18**(10):10－12.