CHEMICAL ENGINEERING TRANSACTIONS

VOL. 71, 2018

A publication of AIDIC

The Italian Association
of Chemical Engineering
Online at www.aidic.it/cet

Guest Editors: Xiantang Zhang, Songrong Qian, Jianmin Xu
Copyright © 2018, AIDIC Servizi S.r.l.
ISBN 978-88-95608-68-6; **ISSN** 2283-9216

DOI: 10.3303/CET1871172

Systematic Analysis of Chemical Dangerous Product Leakage based on Large Data Technology

Qin Zhang [a], Yutang Liu [b]

[a] School of Mathematics and Information Science, Xinxiang University, Xinxiang, Henan 453000, China
[b] Department of Basic Subjects, Henan Institute of Technology, Xinxiang 453002, China
zq821130@126.com

Due to facility upsizing of the chemical system itself, the high process continuity and the complicated inter-parameter mechanism, the chemical system has its own characteristics in addition to the above 4V (Volume, Variety, Velocity, High Value) features: high dimensionality, strong non-linearity, unevenly distributed sample data, low signal to noise ratio. It is due to these unique features of chemical system that there have some difficulties in the analysis and mining of its big data from the traditional methods. In this paper, Chemical characteristics of system is constructed based on big data platform architecture, and uses the hybrid diagnostic identification algorithm, preidentification of abnormal state of chemical system that may arise, to prevent and avoid the effect of practical application. By analyzing the error based on exception identification method of single point parameter, we can see that the method of abnormal identification of chemical systems based on large data technology has high stability and reliability.

1. Introduction

1.1 Overview of Big Data Technology

In 2001, the concept of the "Big Data" first emerged in a Gartner research report. Entered in 2012, the phrase Big Data was increasingly mentioned. The people always use it to describe and define mass data generated in the age of information explosion. Now Big Data has become a synonym for the latest technologies and innovations (Cui, 2016). More and more government, enterprises and other institutions come to realize that data tends to be the most important asset of the organization while the data analytical capacity is no doubt the core competitiveness of the organization.
In general, the Big Data has four key features: Volume, Variety, Velocity, Value, that is, the so-called "4V" feature.

1.1.2 Variety

Data diversification is reflected on two fronts: multiple sources of data, such as SNS, web logs, call records, search engines, etc.; multiple formats of data, such as structured data, semi-structured data and unstructured data.

1.1.3 Velocity

Relevant data statistics show that Taobao "Dual Eleven" activities is surged with tens of millions of traffic every minute, while the PV of 12306 Website on peak day of 2015 spring transportation reaches 29.7 billion with 1032 tickets sold per second, which means the system should enables to process and analyze data efficiently, rapidly, and stably.

1.1.4 High value

In early 2008, Alibaba mined and analyzed data about user behaviors and found that the number of buyers' inquiries swooped, while the China's exports to Europe and the United States declined. On these grounds, they predicted the trend of world foreign trade economy and successfully evaded the financial crisis (Duan et al., 2014).

Please cite this article as: Zhang Q., Liu Y., 2018, Systematic analysis of chemical dangerous product leakage based on large data technology, Chemical Engineering Transactions, 71, 1027-1032 DOI:10.3303/CET1871172

1.1.4 Volume

On March 1, 2013, IDC published its latest digital universe report, which said that as the mobile devices such as PCs and smart terminals grow in popularity, the traffic of SNS hikes up as well as data generated by surveillance equipment and sensors presents an explosive growth, now up to an astonishing figure 2.8ZB, it is predicted that by 2020, the digital universe will reach 40ZB (Gong et al., ,2015). The global data volume is predicted as shown in Figure 1

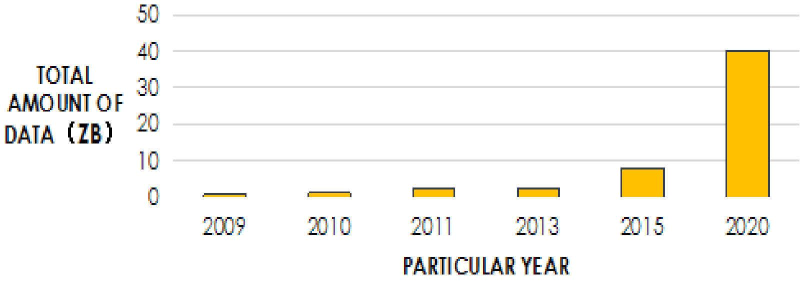

Figure 1: Trend map of total global data increase in different years

1.2 Features of Big Data in Chemical System

Due to facility upsizing of the chemical system itself, the high process continuity and the complicated inter-parameter mechanism, the chemical system has its own characteristics in addition to the above 4V features:

1.2.1 High dimensionality

In the process of chemical industry production, many kinds of physical and chemical changes occur with various parameters highly coupled, thus forming a complex and volatile system. The description of these processes is high dimensional (Tan et al.,1993).

1.2.2 Strong non-linearity

The relationships among various parameters in the chemical industry are non-linear, for example, there is a typical nonlinear relationship among pressure, temperature and entropy and enthalpy in the thermodynamics, and between reaction temperature and velocity.

1.2.3 Unevenly distributed sample data

In the production process, various equipment and parameters should be run in an unideal state, but for various reasons, the parameters will have certain fluctuations.

1.2.4 Low Signal to Noise Ratio (SNR)

Although today's measurement and sensor technologies have advanced to a higher level, there are much noise in data acquired due to some objective factors, such as damage to the instruments and distortion during data signal transmission.

It is due to these unique features of chemical system that there is a certain disparity in the analysis and mining of its big data from the traditional methods. (Gao et al., 2014).

1.3 Unusual classification of chemical system

Chemical system may be interpreted into a process by which one or more products with completely different chemical and physical properties can be made from raw materials or other types products via certain input and output activities (Cheng et al., 2015). There are many types of chemical enterprises in China, involving many industries. The chemical products as abnormal clusters can be divided into the following five categories based on these features, see Table 1.

Table1: Abnormal classification of chemical systems

Abnormal classification	Technological technology	Equipment, raw materials	Human factor	Environmental factor	Other
Percentage (%)	20.17	38.98	7.29	20.17	3.39

In recent years, big data technology has been applied well in finance, trade and health care industries, but the application of big data technology in chemical system is still in its infancy. This paper from the characteristics, big data and chemical system analysis method and application of the 3 aspects are introduced, the process data in addition to 4V has the general characteristics of large mass of data, diversity, speed and variability, but also has the characteristics of high dimension and strong non-linearity, uneven sample distribution and low SNR ratio. Chemical characteristics of system is constructed based on big data platform architecture, and uses the hybrid diagnostic identification algorithm and single point parameter anomaly identification method, pre identification of abnormal state of chemical system that may arise, to prevent and avoid the effect of practical application. It is pointed out that we should analyze and excavate the production data and the market data of raw materials and products in the future, so that we can give greater play to the value of big data.

2. Big data platform architecture of chemical system

Big Data Platform of chemical system underlies the chemical production management carried out with Big Data technology. It provides good data samples and analytical algorithms against the occurrence of exceptions in the production process of chemical system and supports system applications (Tao et al., 2012). The distributed technology architecture can be used to build big data platform to facilitate the rapid data acquisition and analytical algorithm; the module-based development technology is adopted to implement a highly customizable user interface for easy expansion and improve the information presentation mode. Platform is built by data integration and sample data model, analytic algorithm development and model application. The platform technology architecture is shown in Figure 2

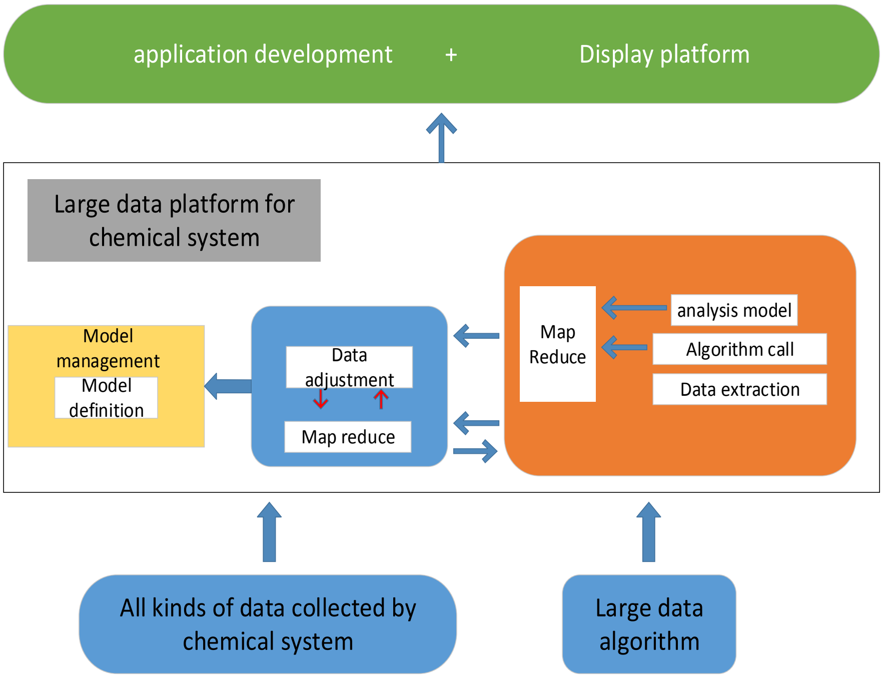

Figure 2: Large data platform architecture for chemical systems

3. Identification measures against exceptions in chemical system

3.1 Hybrid Diagnosis Identification Algorithm

The mathematics model of hybrid diagnosis identification algorithm is given as:

$$Q_l = \frac{\partial Q_l}{\partial x_v} x_v + \frac{\partial Q_l}{\partial p_l} p_l = K_q x_v - K_c p_l \tag{1}$$

Continuity equation is given as:

$$Q_l = A_1 \frac{dy}{dt} + \frac{V_t}{2(1+n^2)\beta_e} \frac{dp_l}{dt} + C_{tc} p_l + C_{tc1} p_s \tag{2}$$

The force balance equation of hybrid diagnosis identification algorithm is given as:

$$A_1 p_1 - A_2 p_2 = A_1 p_l = m_L \frac{d^2 y}{dt^2} + B_c \frac{dy}{dt} + Ky + F_L \tag{3}$$

With Laplace transformation for (1), (2), (3),

$$\begin{cases} Q_I = K_e X_V - K_c p_I \\ Q_I = A_I sY + \dfrac{V_f}{2(1+n^2)\beta_e} sp_I + C_{fc} p_I + C_{fc} p_s \\ A_I p_I = (m_I s^2 + B_c s + K)Y + F_I \end{cases} \tag{4}$$

$$\frac{Y}{X_v} = \frac{\dfrac{K_q}{A_1}}{s(\dfrac{s^2}{\omega_k^2} + \dfrac{2\xi_k}{\omega_k} s + 1)} \tag{5}$$

Where

$$\omega_k = \sqrt{\frac{2(1+n^2)\beta_e A_1^2}{m_L V_t}} \tag{6}$$

$$\xi_k = \frac{\beta_c}{2A_1}\sqrt{\frac{V_t}{2(1+n^2)\beta_e m_L}} + \frac{K_c + C_{tc}}{A_1}\sqrt{\frac{(1+n^2)\beta_e m_{Ltt}}{2V_t}} \tag{7}$$

The transfer function of hybrid diagnosis identification algorithm can be obtained.

Exceptions occurred in the chemical process will be identified with big data technology. First, it is required to establish an exception identification platform, that is, based on the big data platform described in the previous section, data sources for exception identification are thereby provided. On the whole, hybrid diagnosis algorithm is dependent on signal- and data-based treatment and knowledge-based approaches. The specific structure is shown in Figure 3:

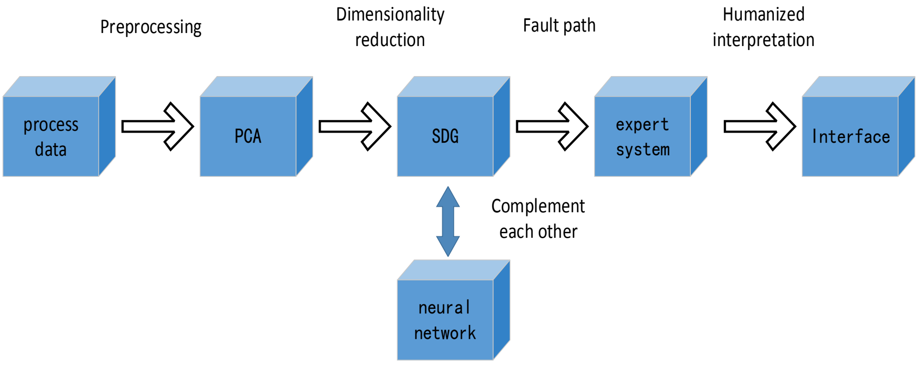

Figure 3: Flow chart of hybrid diagnostic identification algorithm

Data about processes directly access to the PCA module after being filtered and preprocessed. Due to the continuity of the chemical system, there is a certain relationship between many variables. The number of variables can be significantly reduced after a principal component analysis, thus streamlining the SDG inference. It is applicable specially to complex large systems. Whenever process data is delivered, the first step is to detect the alarm if occurred in a chemical process. The PCA is used to analyze these alarm nodes to find the principal component and only analyze its nodes. In this way, the scope of SDG inference can be diminished in order to increase the speed of it. After reducing process data of dimensionality, the number of variables decrease somewhat, thus the well-established SDG model. What counts is streamlined, as the number of nodes in the model decreases, the inference of the SDG model can be greatly accelerated. This is extremely conducive to online fault diagnosis (Luo et al., 2011).

Due to the limitations of human cognition, there are some phenomenons human beings cannot explain, or even explained ambiguously. Especially in the chemical process, it is difficult to express some processes in precise language or by a formula. The causal relationship between some variables is also far more than simple "+" and "-". At this time, the powerful learning function of the neural network comes in for "learning" these faults, since the neural network is a "black box" learning method, no need to know the fault mechanism. Neural network can be trained with changes in data about symptoms the failures manifest and their external symptoms so that it can timely find which problems occurs next time (Pan, 1998).

In the SDG model, a node represents a variable, also does a device, a phenomenon. During SDG modeling, a certain number of nodes are reserved to indicate some devices or phenomena the SDG cannot explain. If these nodes are reached for inference, neural network can be invoked to infer or explain them. In this way, the precision of SDG can be greatly improved, so does the completeness of SDG diagnosis. After getting the abnormal path using SDG, the knowledge base of the expert system is called to make a reasonable explanation for such path, and then display the diagnostic results on the man-machine interface to facilitate the exception handling. In doing so, the process of exception identification is ended (Wang et al., 2015).

3.2 Exception identification method of single point parameter

Exception detection of single measurement point was carried out by using Pauta criterion based on the relevant parameters. The Pauta criterion (3σ) cannot work on smaller sample size, and on the premise that the data sequence to be detected is subjected to normal distribution. The so-called 3σ means that for object data X with normal distribution, the probability that X falls within the interval [e-3σ,e+3σ] is greater than 99.7%, while that it falls outside this interval is less than 0.3%. The data points where this part falls outside the interval [e-3σ,e+3σ] are considered as outliers.

The omission rate and the false drop rate are used to illustrate what is the effect of this method for detecting exception points, the computational formula (8) is given as follows:

$$\begin{aligned}&\text{Leakage rate}=\frac{\text{Number of undetected outliers}}{\text{Total number of data}}\\&\text{noise factor}=\frac{\text{Number of normal points detected by mistake}}{\text{Total number of data}}\end{aligned}\tag{8}$$

For chemical system, it is not certain whether the measurement points must be subjected to the normal distribution, however in the regression analysis of the predilections of parameters, the error distributions of the measured value and the predicted value are suggested to the normal distribution. Due to relatively better robustness of the median than magic number, it is not easy to be impacted by the abnormal point. The robustness of the algorithm can be improved if the median error and absolute deviation are replaced with the original error and standard deviation.

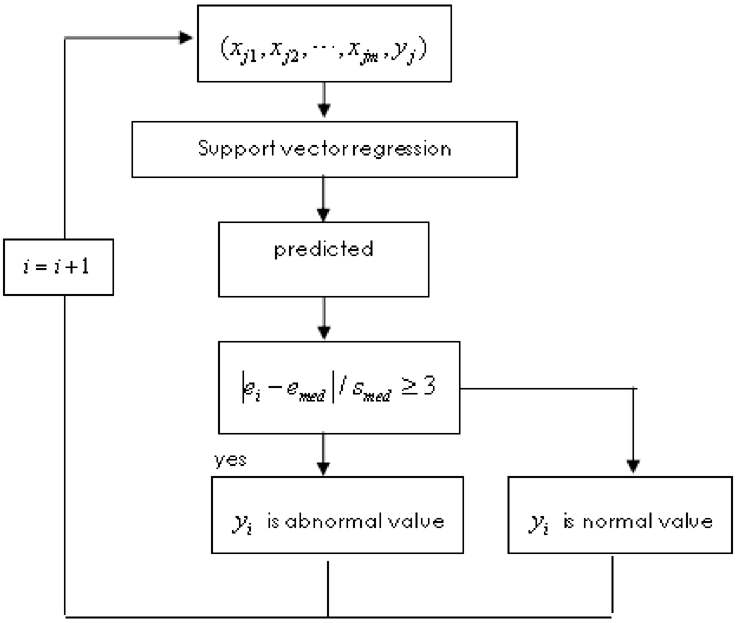
$(x_{j1}, x_{j2}, \cdots, x_{jm}, y_j)$
Support vector regression
predicted
$i = i+1$
$|e_i - e_{med}| / s_{med} \geq 3$
yes
y_i is abnormal value
y_i is normal value

Figure 4: Modified data anomaly detection flow chart

On this basis, this paper uses a modified 3σcriterion, as described below: Suppose the actual value of the point to be detected is y(i=1,2,3,…,n) and m parameters associated to it is x_{i1}, x_{i2},…, x_{im}, a set of samples is obtained. The support vector regression algorithm improved by the genetic one is used to predict y with {(x_{j1}, x_{j2},…, x_{jm}, ,y_j)}as the input pair. The predicted value is $\hat{y}_i$, the difference of which from the actual value is e_i= |y_i-y_{im}|. e_i is sorted in the ascending order, where the error median median(e_i) is imported into correct criteria 3σ. In general, if a data point satisfies |e_i-e_{med}|/S_{med}≥3, that is, the threshold E=3, then the point is regarded as abnormal data. Figure 4 shows the flow chart of data exception inspection based on corrected 3σcriteria. The dual point parameters can be check for abnormalities in accordance with the parameters of a single point, and then the average value is evaluated.

4. Conclusion

In this paper, based on the characteristics of the chemical system construction of big data platform architecture, and uses the hybrid diagnostic identification algorithm and single point parameter anomaly identification method, pre identification of abnormal state of chemical system that may arise, to prevent and avoid the effect of practical application. Good results have been achieved. However, these big data technologies basically stem from the principles and methods of the original data mining technology, and have not fully displayed the characteristics and potential values of big data. That is to say, big data technology has been applied, its essence is the technology of data mining, although the data dimension and the amount of data increases, but also did not make full use of the existing data, to a certain extent, is not conducive to the development and progress of industrial technology. In the future, we should analyze and excavate the production data and raw materials and market data, so that we can give greater play to the value of big data.

Reference

Cheng J., Zhang Z.Y, Ren X., Li S.J., 2015, Prediction of abnormal events in chemicalprocess based on Bayes theory and Vine Copula, Journal of East China University of Science and Technology (NATURAL SCIENCE), 41(2), 144-150, DOI: 10.14135/j.cnki.1006-3080.2015.02.002

Cui L.Y., 2016, Big data break chemical industry park multi dragon flood problem -- an interview with the National People's Congress, Jiangsu Menglan Group Chairman Qian Yuebao, DOI: 10.19474/j.cnki.10-1156/f.2016.07.019

Duan Q.X., Yuan T.J., Mei S.W., Chen J., 2014, Energy coordinated control strategy for wind energy hydrogen storage and coal chemical multi energy coupling system, high voltage technology, 1, 1-11, DOI: 10.13336/j.1003-6520.hve.20171227022

Gao Z.Y., Huo W.H., Gao J.M., Jiang H.G., 2014, Diffusion mapping and anomaly identification of massive data in chemical systems, computer integrated manufacturing system, 20(12), 3091-3096, DOI: 10.13196/j.cims.2014.12.020

Gong Y.H., Xu Y.,2015, Dynamic risk assessment of chemical systems based on dynamic fault tree, security and environmental engineering, 22 (2), 134-138, DOI: 10.13578/j.c-nki.issn.1671-1556.2015.02.026

Luo G.S., Wang K., Wang Y.J., Lv Y.C., Xu J.H., 2011, The principle and application of microchemical system, chemical progress, 30 (8), 1637-1642, DOI: 10.16085/j.issn.1000-6613.2011.08.014

Pan Y.,1998, Modern chemical industry in the application of chemical system optimization technology, (4): 31-32, DOI: 10.16606/j.cnki.issn0253-4320.1998.04.013

Tan X.F., Li P.,1993, Selection of abnormal data in chemical experiments. Shenyang chemical, (3), 17-18+52, DOI: 10.13840/j.cnki.cn21-1457/tq.1993.03.006

Tao S.H., Li C.K., 2012, Variable abnormal sequence method for fault diagnosis in chemical process, computer and applied chemistry, 29 (2), 178-180, DOI: 10.16866/j.com.app.chem2012.02.013

Wang D.W., Sun Z.W.,2015, Power user side big data analysis and parallel load forecasting, proceedings of the Chinese Academy of electrical engineering, 35 (3), 527-537, DOI:10.13334/j.0258-8013.pcsee.2015.03.004