Research on Environmental Monitoring System Based on Microservices and Data Mining

Yu Liu, Junge Huang*, and Ningqi Lu

College of Urban Construction and Safety Engineering, Shanghai Institute of Technology, 201418 Shanghai, P.R.China

Abstract. The monitoring systems of various industries have various types and different structures. There are problems of “data chimney” and “information islands”^[1]. Monitoring data is difficult to be effectively utilized and cannot provide reliable data information to support for environmental security^[2]. In this end, an environment monitoring system based on micro-service architecture is designed. The information management and automatic monitoring business systems are unified into a flexible, robust and efficient system platform to adapt to the big data analysis and the mining applications. Using Hadoop to build environment monitoring big data platform, distributed storage, selective extraction and efficient calculation of the massive environment monitoring data can be achieved. By integrating the detection and monitoring data of the ecological environment and in-depth mining it, a neural network model is established to automatically identify potential safety hazards and recommend corresponding treatment measures, so to assist in the comprehensive research and scientific decision-making of environmental safety and promote intelligent management of safety.

1 Introduction

Under the craze of the Internet, many enterprises have upgraded and integrated current information systems to move towards intelligentization. Environmental monitoring field is no exception. There are many kinds of existing business systems related to environmental monitoring, and they are different and independent. With the individualization of the requirements of emerging systems and the shortened of its product life cycle^[3], it is inevitable to choose a flexible, expand easily, scalable and highly available system architecture. The current popular micro-service architecture is able to build a flexible, robust and efficient system platform which integrates with many small services. This system platform integrates information management, automatic monitoring and other business systems to solve the problems of independent environmental monitoring systems that is low utilization of data information. The designed big data platform utilizes massive monitoring data for analysis and mining, and introduces machine learning algorithms into the system, which uses artificial neural network algorithms to predict, classify and identify data, and recommends corresponding processing measures to promote research on the intelligent management of the environmental safety. In the end, an ideal intelligent integrated system is provided, which includes functions such as information management, business processing and data analysis modeling and

automatic identification, alarm and recommending processing measures.

2 System Design Based on Micro Service Architecture

The goal of the microservice architecture is to split a complex application into multiple microservice modules. Each module provides a single service function to the outside, and the coupling between modules is reduced. Each module uses its own framework and development language. They do not interfere with each other and can be independently compiled, deployed, and run. At the same time, each module communicates with each other through standard interfaces, and combines as a whole to provide complete services to the outside^[4]. Since each module is deployed independently, each has a memory space that does not interfere with each other, and modules cannot be directly called. Therefore, a message gateway is required to provide communication access services for all microservices.

Combined with the characteristics of micro-services, the original monitoring, inspection and reporting services and some common information system modules are divided into a group of small services. The services communicate through the HTTP-based RESTful style API or the message driver as the calling mode. The multiple sub-blocks included in the system constitute the micro-service architecture system together. The infrastructure of the system is shown in Figure 1.

* Corresponding author: Junge Huang. Email: hjg@sit.edu.cn

Fig. 1. System infrastructure.

The system is divided into three levels: the front end, the business layer, and the data platform. The front end is mainly that the user interacts with the service gateway through a specified access address to interact with the service layer. The service layer actually provides and implements various service interfaces, wherein it includes microservices registry and messaging service middleware. Which is the storage, analysis and mining of data, the data platform includes database sets, big data platforms, and intelligent modules such as machine learning and alarm. The front-end and business layer are mainly based on the Spring Cloud implementation, providing services such as the service registration, the service discovery, and load balancing, which allows the system to have more computing resources when scaling horizontally. One of the servers is heavily loaded, and the new requirements are forwarded to other idle machines. When a service fails, the blowing mechanism is triggered Hystrix, returns a result identification error to the service caller, releasing the thread in time to prevent the failure from spreading in the distributed system. Each module uses the message middleware to communicate data by using an efficient and reliable message delivery mechanism, and connects the business layer with the data platform to realize communication between different platforms.

3 Construction of Environmental Monitoring Big Data Platform

After the integration of environment detection and monitoring data, deep mining based on machine learning needs to be implemented on the big data platform. The platform mainly uses Hadoop's powerful data management tool - HDFS and fast and simple MapReduce statistical tool - Hive^[5] to construct the data warehouse by SQL-like and analyze various monitoring information. Hadoop-based data acquisition system - Flume^[6] is used to collect auxiliary data such as alarm information and system logs in real time. Finally, by the message service middleware, it is associated with the database cluster, the business system and the machine learning module to realize the timely display and access of the analysis data, so that the machine learning can not only analyze the model offline, but also predict the recommendation online, forming a message-driven machine learning service^[7]. Machine learning is mainly divided into two modules: modeling and prediction. The

modeling module needs to learn a large amount of historical data to build a model. It requires direct access to the data source for analysis and learning, which does not consume each data through the message middleware. The machine learning online prediction needs to listen to each message received by the message middleware to consume and process them. It uses the machine learning algorithm to predict, and feed back the prediction result to the message middleware for use by other services. The message-driven machine learning architecture is shown in Figure 2.

Fig. 2. Message-driven machine learning service.

Proceed as follows:

- (1) The message source sends the message to the message service middleware, and simultaneously copies a copy to the machine learning offline modeling module. The module also collects the data through the big data platform extraction and real-time monitoring, and then learns and builds the model.
- (2) After the machine learning offline modeling module receives the message, it conducts historical data learning modeling. There are two main methods of modeling, modeling of tag data and modeling of tagless data. The data of the existing tags can be implemented by supervised learning algorithms such as Random Forest algorithm and Bayesian algorithm. The modeling of unlabeled data needs to be realized by unsupervised learning algorithm, mainly K-means algorithm, and PCA backbone analysis method is used to realize cluster analysis and outlier sample acquisition.
- (3) The machine learning online prediction module listens to the new message TOPIC of the message middleware. After performing the same preprocessing as the offline modeling, the execution algorithm obtains the prediction result, and then the prediction result is written into the message middleware in real time for other microservices to execute forecast result.

4 Identification and Handling of Safety Hazards

4.1. Data exploration and preprocessing

The data in the environmental monitoring system is provided by sensors and manual inspection. The system needs to collect data before the machine learning module is modeled. These data include the real-time monitoring data of each business system, and collected data by the Hadoop-based data acquisition system Flume from the

system database on time. Second, the system uses the data warehouse tool, Hive, to clean and count data on the Hadoop platform. Data cleaning requires the establishment of an external table in the library. After the cleaning is completed, the structure is stored in the partition table to speed up the statistical analysis. The data statistics are mainly carried out by HiveQL. The execution process of this part is: submit the HiveQL to the Hive-Server2 service through the JDBC interface, complete the query plan through interpretation and compilation, and then process the completed plan through the executor, and finally call and execute by MapReduce. The pre-processed modeled sample data is extracted and awaiting training. The data preprocessing model flow is shown in Figure 3.

Fig. 3. Data preprocessing model flow.

(1) Data integration, which integrates raw data in

chronological order.

(2) Data protocol, due to the large number of recorded data attributes of the data acquisition device, the model makes the following data specifications for the modeling data:

Attribute conventions, attributes that have no effect on model construction can be removed;

The numerical specification can be stipulated for some data records whose value is 0 and whose properties are unchanged.

(3) Data transformation, this step mainly performs attribute construction according to the established data table model to facilitate data cleaning.

(4) Data cleaning, in addition to invalid data processed in the data protocol, further deletion or addition is required.

4.2. Construction of Neural Network Model

Here, the gas concentration prediction and over-limit identification of the coal cutting face in the ventilation monitoring system is taken as an example. Using the RBF neural network^[8] algorithm and a large amount of historical data for training, the system sets the sample data of the previous shift, and the time and gas concentration as the first layer (input layer) attribute. Set $x_i \in R^n (i = 1, 2, \dots, n)$ as the input of the RBF neural network, $y \in R$ is the output of the RBF neural network, as shown in Figure 4.

Fig. 4. RBF neural network.

Set

$$y = f(x) = \sum_{i=1}^n \omega_i g_i(x) + b \quad (1)$$

Among $g_i(x)$ is the function of the hidden layer - the radial basis function. Let the radial basis function be

$$g_i(x) = \exp\left(-\frac{\|x - \mu_i\|^2}{\sigma_i}\right) \quad (2)$$

μ_i is the center of the radial basis function; σ_i is the width of the radial basis function.

During training, the weight between the hidden layer and the output layer can be directly calculated by the least square method. It can be determined by learning, or the value specified by the system for the gas prediction can be selected. The parameters are optimized in the system, and the neural network with two hidden layers is found to have better training effect. Therefore, the next shift and the next day shift of the corresponding time can be compared and tested by the model to more accurately predict the gas concentration in the subsequent period. If it is predicted that the gas concentration is about to exceed the limit, it can enter the safety identification model, identify the impending safety accident, and recommend the corresponding treatment measures.

The BP neural network model is selected for the security identification, which is a classic algorithm in neural network learning. The BP neural network algorithm is very similar to the RBF, but the RBF is better than the BP. Because RBF shows a stronger advantage in the fields of nonlinear function approximation, time series analysis, data classification, etc. It can perform a wide range of data fusion and process data in parallel at high speed. In the security identification, the corner of the coal cutting face will soon exceed the limit. The recommended method of module handling is to "set the windshield and clear the accumulated gas". If the position is continuous or exceeds the maximum threshold, the recommended treatment will be "stop cutting coal".

In order to establish the best neural network model, it is mainly analyzed from the following four points:

(1) Input layer and output layer neuron selection.

First, determine the number of neurons in the input layer and output layer of the neural network. Next, repeatedly train through the network according to the actual collected and saved data. At last, use the previously collected data to predict the data of the next stage.

(2) Hidden layer design.

According to the actual situation and the actual collected data, the input layer and the output layer are set. Under considering the network performance, the number of neurons in the network hidden layer is set.

(3) Transfer function selection.

The transfer function plays a very important role in the neural network. Such as Tansig function and Logsig function in the BP neural network.

(4) Training function selection.

The commonly used training functions provided by the neural network toolbox have been repeatedly trained. For example, using the gradient descent method, the L-M optimization algorithm performs weight training.

4.3. Realize of Timely Alarm

The timely alarm of the system is actually based on the prediction module. In the system, let the client establish a long connection with the server through WebSocket. As long as one user is online, the server will open the thread that polls the alarm information. Once it finds that the alarm information is written into the database, it immediately sends an alarm message to all online related users. The server only needs to open a thread to implement the alarm. It does not need to poll every user, which reduces the pressure on the server. The alarm information is generated by the gas monitoring information table of the trigger monitoring database, and when the data is inserted, the over-limit judgment is performed, and the over-limit data is classified and then written into the alarm information table.

5 Conclusion

The research shows that the micro-service architecture can be used to realize the unified integration of various business systems of the environmental safety monitoring system, and solve the problems of “data chimney” and “information island”. The API-based architecture form of micro-services is able to solve the interaction problem between the business layer and the front-end. The message-driven architecture is able to solve the data interaction between the business layer and different platforms. The system can also use the message middleware to realize offline learning and online prediction. Among them, the message-driven micro-service architecture combines the usability of the peer-to-peer architecture with the reliability of the API-based gateway architecture. It is a relatively balanced micro-service architecture solution, because the message middleware caches all messages until the confirmation message is sent or received, which ensures the reliability of message publishing and consumption, which in turn increases the reliability of the entire microservices architecture. The monitoring software system can scientifically and effectively manage the ecological environment, timely monitor the surrounding environment and assist in the pre-judgment of safety accidents, etc., which facilitates monitoring and testing, reduces maintenance costs, and ensures safety. It is an

important system engineering. However, for the research and experiment of this system architecture level, it is necessary to compare and analyze, improve the system operation mode, and combine the machine learning module and business to conduct in-depth research to realize the full intelligence of environmental monitoring.

Acknowledgment

Foundation item: National natural science foundation of China (41564005)

Author Introduction: Yu Liu(1995-), Male, studying for master's degree, research direction: Computer technology and security engineering. Email: liuyu9510@hotmail.com
Corresponding author: Junge Huang(1966-), Male, postdoctoral, professor, research direction: Civil engineering detection and monitoring, Safety engineering. Email: hjg@sit.edu.cn

References

1. B Li. Analysis and design of coal mine safety monitoring system. Master's Theses. Beijing University of Posts and Telecommunications(2009).
2. Y.Z Cui, M.L Bai, B Li. Research on Key Technologies and Development of Intelligent Mine Big Data. Journal of Coal Science and Technology. 47(3),66-74(2019).
3. P Gong. Microservices distributed architecture development combat. People's Post and Telecommunications Press(2018).
4. G.R Ruan. Preliminary Study on the Development of Digital Platform for Construction Enterprises. Journal of Building Construction. 4: 694-696(2019).
5. L.J Zhang, Z Fan, W.C Wei. Hadoop and Big Data Mining. Mechanical Industry Press(2017).
6. Xipu Education Research Institute, Y.T Shi, X.G Lin. Hadoop big data development case tutorial and project combat. People's Posts and Telecommunications Press(2017).
7. C.W Zhao. Analysis of message-driven machine learning microservice architecture. Journal of Computer programming skills and maintenance. 24: 68-69(2017).
8. H.Z Yu, L.Z Xu, H.B Wang. Monitoring and Monitoring Information Fusion Technology. Tsinghua University Press(2011).