Scientific Research Publishing

Journal of Computer and Communications, 2019, 7, 1-7
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

# Research on the Anomaly Detection Method in Intelligent Patrol Based on Big Data Analysis

**Xiaoqing Deng**

School of Intelligent Manufacturing, Sichuan University of Arts and Science, Dazhou, China
Email: 47872614@qq.com

**How to cite this paper:** Deng, X.Q. (2019) Research on the Anomaly Detection Method in Intelligent Patrol Based on Big Data Analysis. *Journal of Computer and Communications*, **7**, 1-7.
https://doi.org/10.4236/jcc.2019.78001

**Received:** August 1, 2019
**Accepted:** August 18, 2019
**Published:** August 21, 2019

## Abstract

The network anomaly detection in intelligent patrol is based on the trigger of a single threshold of network element performance parameters in patrol task, which has a high false alarm rate and low efficiency. In order to effectively and accurately integrate network performance, this paper proposes to mine network element performance data and network element log information in the integrated automatic patrol to detect network anomalies. Because log files have a large amount of data and a variety of types, and log data has a complex structure and contains large implied information. The relationship between network anomalies and time can actively discover through the analysis of the log files. Therefore, big data mining and classification can greatly improve the efficiency of data processing. However, the accuracy of finding network anomalies is insufficient only for log analysis. Therefore, this paper puts forward the performance indexes collected in the log analysis and patrol inspection system and adopts the sequence analysis algorithm to detect network anomalies, so as to improve the accuracy and efficiency of detection.

## Keywords

Big Data, Intelligent Patrol, Anomaly Detection

## 1. Introduction

With the emergence of cloud computing, integration of three networks, Internet of Things and mobile internet, the massive growth of data and the ever-changing types of data indicate that we have entered the era of big data. The use of big data technology has increasingly become a powerful weapon for enterprises to go beyond their competitors. How to ensure the effective, stable and safe operation of enterprise network has become the key to enterprise operation. How to find problems efficiently and timely, eliminate hidden dangers and prevent accidents has become a key problem to be solved in different industries. In order to meet the needs of daily network patrol work, a large number of intelligent electronic patrol systems have emerged in a large number of enterprise network management systems. With the rapid development of network technology, patrol work is to maintain the safe and stable operation of network of all walks of life, and the safety and stability of the industry are very important, so the network anomaly detection is necessary. The traditional patrol method is mainly manual patrol, supplemented by other methods. Failures and hidden dangers are generally reported by telephone, and the patrol records are mainly managed by hand, so the traditional patrol method lacks the real-time tracking and monitoring of patrol, and quantitative assessment of patrol maintenance quality. The automatic reporting and processing of patrol results and hidden troubles cannot be realized. The efficiency of patrol management is relatively low. People have begun to consider introducing artificial intelligence into the patrol system, constructing the intelligent patrol system, and using the machine learning method to predict network anomalies. This paper proposes to complete mining of log data and extract features through the big data analysis technology, comprehensively analyze network element performance data, use the time series analysis algorithm to discover network anomalies, and improve the network security performance.

## 2. Outline of Anomaly Detection

Network anomaly refers to the change of data inflow and outflow beyond a certain range due to the gradual exhaustion of system resources (CPU load, memory usage, etc.) or network attacks (abnormal processes), resulting in system crash or user’s inability to work. Anomaly detection refers to identifying and discovering behavior data objects that do not meet expectations through a certain method so that the administrator can manage the process and adjust the configuration of network resources [1]. In the field of anomaly detection, there are a lot of studies. At present, a large number of network anomaly detections are based on setting threshold value for a single performance index parameter of network elements to detect and judge specific network equipments, and give an alarm when a failure occurs. However, this method has strong dependence on indicators, and the accuracy of performance indicators directly affects the efficiency of failure detection [2]. Moreover, the performance indicators contain less information, which can fully reflect the specific operation status of the equipment, and have a low coverage of failure types. Literature [3] proposes that the classification is made based on network traffic with decision tree to make anomaly prediction. Literature [4] proposes that multiple variables are predicted by Markov chain based on network traffic, and the predicted values are classified by Bayesian classification to make anomaly alarm. These models can only detect anomalies according to the existing classification. The historical data needs to be classified manually, and only the changes of values are considered. There is a big error in the prediction results.

In the process of network operation, log data records the state of network, user's operation, data changes and so on. It is closely related to network anomalies. Network anomalies and resource changes can be effectively detected by analyzing log data. Therefore, the analysis of system log provides a powerful basis for the accurate judgment of network anomaly detection. In the large-scale network environment, the frequency of log generation by network element devices is at the minute-second level, and the delay of log data from generation to analysis can be controlled at least at minute level to ensure the accuracy and timeliness of early warning messages. The computing method based on batch processing cannot meet the requirements of computing one by one. In addition, the network equipment logs contain a lot of redundant log (periodic notification logs, irrelevant concurrent logs caused by failure, etc.) besides failure information, so that valuable failure information is submerged in a large amount of invalid information. It is difficult for the traditional abnormal point detection method to output ideal analysis results. With the development of big data technology, there are relatively good solutions to the difficulties in equipment operation and maintenance scenarios. Literature [5] proposes a network anomaly detection system which combines with the big data technology. The system uses the big data technology to extract a large amount of security data in the network, uses distributed storage to improve the efficiency of data query and processing, and predicts the security trend and risks of events that have occurred and those which are happening. Literatures [6] [7] propose a network anomaly detection method based on multi-scale principal component analysis, which considers the temporal and spatial correlation of traffic matrix, and proposes a whole network anomaly detection method based on MSPCA, and meets the needs of real-time monitoring. Most of the above work focuses on network anomaly detection for network traffic, while little work focuses on comprehensive analysis for anomaly log. In order to improve the accuracy and efficiency of network anomaly detection, this paper proposes a method of combining comprehensive network element performance index with system log data [8]. The abnormal data is accurately detected by using the comprehensive time series analysis method. The comprehensive time series analysis method is mainly based on the difference of log data performance before and after the failure in network logs and combined with the changes of comprehensive performance indicators, which can accurately detect network anomalies and improve the accuracy of anomaly detection.

## 3. Anomaly Detection Method

In the network anomaly detection based on big data, because there is a large amount of network element performance data collected by the patrol system, and a large amount and many types of log information of network element, it is necessary to use the big data technology to realize the structured processing of log data, use the comprehensive time series analysis method by combining with performance data, complete the detection of abnormal data, realize the intelligent early warning of abnormal data, and record the abnormal data in the log.

Figure 1 shows Frame diagram of anomaly detection

### 3.1. Data Acquisition

This paper realizes the accurate detection of abnormal data in the intelligent patrol system. The acquisition of two types of data is mainly included. One is the performance data of network elements acquired by the system, and the other is log information of the system.

The performance data of network elements mainly include the number of performance indicators collected, *i.e.* 1000 - 1500 performance indicators collected within the acquisition cycle of a single acquisition machine. If N acquisition machines are deployed, 1500 × N performance indicators will be automatically returned. The time granularity is acquired. If the long-term monitoring is required in daily maintenance, the minimum acquisition time granularity of the indicators is 5 minutes. The minimum time granularity of instant monitoring indicators of a small number of hotspot polling is 5 seconds.

The front-end function of automatic patrol receives the patrol task instruction files issued by the application management layer through the northward interface of the data acquisition layer. The patrol task instruction file is an XML file. Its file naming rule: unique identification of network management server + patrol task identification.xml, according to following form Table 1.

Log acquisition content varies by the system served by network element equipment, network and host equipment. The main objects of log acquisition include Unix host system, database, firewall, network equipment (router, switch). The log acquisition module acquires logs through related equipment and business system defined in data acquisition strategy. The data of the acquisition server is uploaded to the network management server. In order to reduce the amount of data, the larger log files are compressed. The naming method of log data file: "alias of network element-acquisition object identification-type of log-acquisition timestamp". The related fields in the name are taken from the associated log acquisition strategy. After the log data is acquired, the acquired results are uploaded to the network management system through the northward interface of the data acquisition layer. If there are problems in the process of uploading, it is necessary to save the log data that has not been uploaded properly on the acquisition machine until it is uploaded to the network management server properly.

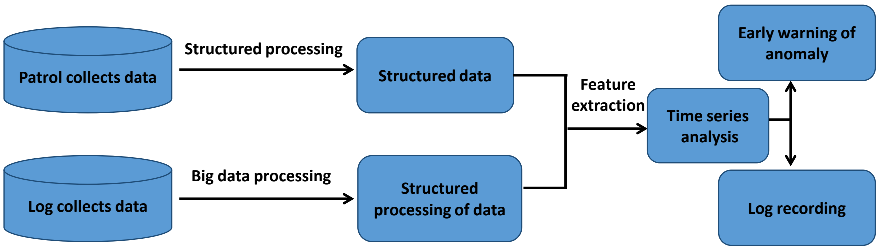

Figure 1. Frame diagram of anomaly detection.

Table 1. Content of patrol task file.

| Field | Field Description |
|---|---|
| Policy ID | Policy ID associated with the instruction file. |
| Interval | Cycle of patrol task (in minutes). |
| FtpSvr Name | The original result file of patrol is uploaded to FTP server. |
| utFile Path | FTP upload path of original result file of patrol. |
| Node Name | List of unique signs of the equipment inspected which uses the patrol instruction. |
| Instruction | Patrol instruction sequence. The execution order is determined by the Execute Order attribute of the element, and the instruction with small serial number runs first. |

Because of the large amount of log data, the traditional log analysis method based on relational database cannot meet the requirements of massive data processing. The big data technology is a powerful means to solve the current problem of log analysis. Based on the big data technology, Literature [9] establishes a parallel network log analysis engine based on Hadoop open-source framework. Under Map Reduce model, IP statistical algorithm and anomaly detection algorithm can effectively solve the problem of a large amount of data and have good performance.

### 3.2. Time Series Analysis Algorithm

The intelligent patrol system is mainly aimed at real-time and continuous monitoring of the network. The data in the network has temporal orderliness and changes with the cycle. Therefore, time dimension is predicted and analyzed as anomaly. In recent years, anomaly researches based on time series mainly focus on point anomaly, pattern anomaly and sequence anomaly. The judgment of point anomaly is the basis of the research of anomaly detection algorithm. After a long period of accumulation, the scientific community has formed a relatively mature time series analysis method. The way to solve the problem is to select a specific prediction method to evaluate the time series according to the characteristics of different time series and to judge whether there is anomaly at this time according to the difference between the predicted value and the actual value. Time series analysis algorithm is mainly used to transform time series data into feature data by extracting feature data of performance and carrying out model matching, etc., and then classify the data. This method has little impact on noise. At present, the research methods of time series are becoming mature. Since the network data in this paper mainly change with time and have time differences, the research method of deterministic time series is adopted [10] [11].

Let the current time be $t$, $y_t$ be the current monitoring value and $p_t$ be the predicted value calculated by the time series method. It is mainly determined by the continuous time smoothing factor $C_t$ and the periodic factor $S_t$. The continuous time smoothing factor is the influence of past value on current time, and the periodic smoothing factor is the influence of monitoring value at the same time in history on current value, reflecting the periodicity of time.

Define the continuous time smoothing factor: $C_t = \alpha y_t + \left(1-\alpha\right)C_{t-1}$;

Periodic time smoothing factor: $S_t = \overline{Y}_t = \sum_{t=1}^{n} Y_t$;

Obtain the predicted value $p_t = \beta C_t + \left(1-\beta\right)S_t$.

Calculate the error between the predicted value and the detected value: $\delta_t = \left|p_t - Y_t\right|$.

According to the error between the predicted value and the actual value, the anomaly of the network is determined. The time series analysis method mainly considers the similarity between recent observed values and predicted values and the periodic characteristic of time series. It can not only detect the non-periodic anomalies in the distribution of detected data in a timely manner, but also discover the anomalies which suddenly appear in the recent time series distribution.

The specific flow of detection with the time series method:

Step 1: Define outliers by analyzing the time series. Mainly analyze the performance parameters and log feature data of the network elements as objects. Locate outliers according to utilization rate of CPU in network elements, utilization rate of memory, traffic, link random loss, link jitter, etc.

Step 2: Through outliers located by Step 1, judge the change of traffic in the ports of network element nodes, whether there are periodic peaks and activestate.

Step 3: Locate outliers by the first two steps, analyze time series data, predict the latest network trend and provide early warning for network security.

The time series analysis method is based on the anomaly detection of network element logs. Combined with the performance data in the analysis patrol system, it can detect anomalies in data, predict the recent network state and give an early warning.

## 4. Conclusion

The development of big data technology has provided strong support to large-scale data mining and massive data analysis in the intelligent patrol system. The log files contain a large amount of information including the state of the network, changes of data and traffic, and user's operation in the process of network operation. The single method is inefficient in processing log data. In order to improve the processing speed of data, big data technology provides good support for this. This paper proposes the analysis and feature extraction of performance data and log data based on big data technology, and network anomaly detection with the time series analysis method. It mainly finds outliers in the time series through log information and performance parameters of network elements, which improves the accuracy of detection. Because of the real-time monitoring of the patrol system, the use of big data technology can quickly realize data analysis and improve the detection efficiency. The next step will be more detailed analysis, data mining and feature extraction, the verification of the time series algorithm through experiments, and the further improvement of the detection accuracy and efficiency.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Casas, P., Vaton, S., Fillatre, L., *et al.* (2010) Optimal Volume Anomaly Detection and Isolation in Large-Scale IP Networks Using Coarse-Grained Measurements. *Computer Networks*, **54**, 1750-1766. https://doi.org/10.1016/j.comnet.2010.01.013

[2] Li, K.T. (2015) Design and Implementation of the Performance Analysis and Early Warning System for the Mobile Core Network.

[3] Gu, X., Papadimitriou, S., Yu, P.S. and Chang, S.-P. (2008) Toward Predictive Failure Management for Distributed Stream Processing Systems. *Proceedings of ICDCS*, Beijing, 17-20 June 2008, 825-832. https://doi.org/10.1109/ICDCS.2008.34

[4] Tan, Y., Gu, X. and Wang, H. (2010) Adaptive System Anomaly Prediction for Large-Scale Hosting Infrastructure. *Proceedings of PODC*, Zurich, 25-28 July 2010, 173-182. https://doi.org/10.1145/1835698.1835741

[5] Hu, T.T. (2014) Research on the Outlier Detection Algorithm in Data Mining. Xiamen University, Xiamen.

[6] Wang, X., Wang, X.L., Ma, Y., *et al.* (2015) A Fast MST-Inspired kNN-Based Outlier Detection Method. *Information Systems*, **48**, 89-112. https://doi.org/10.1016/j.is.2014.09.002

[7] Ying, Y., Ren, K. and Liu, Y.J. (2018) The Network Log Analysis Technology Based on Big Data. *Computer Science*, **45**, 363-365.

[8] Sandford, P.J., Parish, D.J. and Sandford, J.M. (2006) Detecting Security Threats in the Network Core Using Data Mining Techniques. *Network Operations & Management Symposium*, Vancouver, 3-7 April 2006, 1-4. https://doi.org/10.1109/NOMS.2006.1687640

[9] Bai, Y.L. (2011) Research and Implementation of Data Mining Algorithm Based on Hadoop. Beijing University of Posts and Telecommunications, Beijing.

[10] Min, C., Dong-Nian, C., Jian-Hui, Z., *et al.* (2009) Network Traffic Abnormality Detection Algorithm Based on Self-Adaptive Threshold. *Computer Engineering*, **35**, 164.

[11] Hu, B., Chen, Y. and Keogh, E. (2013) Times Series Classification under More Realistic Assumptions. *Proceedings of* 13*th SIAM Conference on Data Mining*, Austin, 2-4 May 2013, 578-586. https://doi.org/10.1137/1.9781611972832.64