Research on Fault Warning of Power Big Data Log Analysis Based on Integrated Prediction Algorithm

Dewen Wang, Yang Li [a]

School of North China Electric Power University, Baoding, 071003, China

[a]7200088@163.com

Abstract

With the completion and in-depth application of various information systems of the State Grid, it provides an important guarantee for the safe and stable operation of the power system. The power information system generates a large amount of log data during the running process, and the traditional log processing system cannot cope with the application analysis of the big data level. Aiming at this situation, this paper proposes a log collection method based on big data platform and uses integrated learning algorithm for data analysis.

Keywords

Big data,log analysis,Fault warning.

1. Introduction

With the continuous expansion of the system scale, a large amount of log data has accumulated in the power information system, which has the characteristics of full-service scope, full-time type, and full-time dimension, and contains key information for the operation of the power system. Real-time collection and analysis of the logs of the information system can timely discover abnormal information or behaviors in the system, which can reduce the troubleshooting time and service interruption time and avoid more serious consequences.

Multi-task distributed technology is used to analyze and mine massive logs. Applying multiple classification algorithms in machine learning, a scientific analysis model can be established, which makes the analysis depth of the log and the recognition accuracy of the event further improved. The significance of log analysis and early warning is that it can explore potential risks in advance, analyze, judge and form qualitative or quantitative descriptions, so as to take countermeasures to reduce risks. This has important theoretical and practical significance for improving the security, stability and service capabilities of information communication systems.

2. Related Work

At present, the grid information system is equipped with relatively complete security protection measures to defend against conventional attacks. However, for the current APT infiltration and other attacks, as well as malicious behavior of internal personnel, there is a lack of quick and timely means of discovery, only through the operation log later. The audit found that low efficiency and long time can only be used for after-the-fact accountability, and it is impossible to find anomalies and block them at the initial stage. Utilize massive user behavior log analysis technology

Surgery can realize timely detection and alarm of abnormal behavior of the user. The management personnel can quickly confirm the abnormality and take timely measures to effectively avoid the serious consequences.

In recent years, relevant research results have been made in the risk assessment and reliability analysis of traditional networks or power information communication networks. However, there are few researches on early warning and early warning of power information system log. Because network risk assessment and early warning brought about by the coupling and correlation of power information and physical system are a new type of problem, further research is needed. For example, the literature [2] puts forward the idea and framework of establishing power CPS, and summarizes the new characteristics of information physical coupling in the risk assessment of power information communication networks. The literature [3] points out that in the context of information physical fusion, due to the large number of transmissions Sensing and monitoring equipment access, so that the sampling data size and sampling frequency are multiplied, greatly increasing the burden of the communication network; the literature [4] pointed out that if the power information network uses a combination of a dedicated network and a general network, then Open communication protocols will bring a lot of potential risks. In terms of risk situation estimation of communication networks, the literature [5] uses cellular automata theory to model and analyze the cross-space propagation mechanism of risk; the literature [6] proposes a topology that reflects the power-communication composite system. The matrix model with intricate relevance has certain advantages in real-time vulnerability assessment; the literature [7] based on complex network theory, analyzes the topology of power communication network from random and deliberate from the perspective of connectivity and network efficiency. The vulnerability under attack, and the high-number node protection strategy and the low-degree node edge-adding strategy are proposed. This method is more suitable for multi-service and large-scale communication networks. [8] Learning rate adaptive BP neural network based on dichotomy The network algorithm performs risk assessment on the power communication network, and abstracts the communication network into a 3-layer feedforward BP network for analysis, which shortens the convergence time to some extent, but may have local optimal deviation and hidden layer nodes. The number is not easy to determine;

3. log analysis method

3.1 Log collection

Flume is a distributed, reliable, and highly available log collection system for massive log collection, aggregation, and transmission. Supports the customization of various data senders in the log system for data collection; at the same time, Flume provides the ability to easily process data and write to various data recipients (customizable). The core of Flume is to collect data from the data source and then to the destination. In order to ensure that the delivery is successful, the data will be cached before being sent to the destination. After the data actually arrives at the destination, the data cached by itself is deleted.

Kafka is a distributed, partitionable, replicable messaging system that maintains message queues. The architecture is very simple and is an explicit distributed architecture. Producer, consumer implements the Kafka registered interface, data is sent from the Producer to the Broker, and the Broker assumes an intermediate cache and distribution. The Broker distributes the Consumer registered to the system. Brokers act like caches, which are caches between active data and offline processing systems. Client-to- server communication is based on a simple, high-performance, and programming- independent TCP protocol.

In terms of log collection, Flume is suitable for configuration solutions without programming. Due to the rich source, channel, and sink implementations, the introduction of various data sources is only a configuration change. Kafka is suitable for solutions that have high throughput and availability requirements for data pipelines. Basically, programming is required to achieve data production and consumption. Therefore, Flume can be used as the producer of data, so that the introduction of the data source can be realized without programming, and Kafka Sink is used as the data consumer, so that high throughput and reliability can be obtained.

3.2 Data processing

The big data stream computing processing mainly includes two high-performance parallel computing engines, storm and spark streaming, which belong to the popular real-time stream computing framework. Spark streaming is a real-time stream computing framework built on spark. It uses the time batch window to generate the calculation input source RDD of spark. Then the job is generated for the RDD, and the queue is scheduled to be executed in the spark computing framework. The bottom layer is based on spark resource scheduling and tasks. Computing framework; Sparkstreaming is a data-based batch processing method, which is calculated for data forming tasks. It is mobile computing without moving data. On the contrary, Storm is in the processing architecture, data flows into the computing node, and the data is moved. Instead of calculations, batch data processing for time windows requires the user to implement it. At the same time, Spark streaming is based on spark, and can be combined with other components of spark to realize interactive query, machine learning MLib and so on. Relatively speaking, Storm is just a streaming computing framework that lacks the integration of the existing Hadoop ecosystem. Spark streaming is more fault-tolerant. Although the real-time performance is slightly weaker and storm, it has little effect on log processing.

3.3 Data analysis method

The algorithm is integrated by multiple classification algorithms in machine learning, and the collected log data is classified and predicted, and then the most accurate optimal solution is found to predict the system fault. Two modes are used for verification. One is to divide the data set into two parts: the training set and the test set, and implement the fault warning function through training test. The second verification mode is to implement the prediction function through the time window based warning log. Using a two-level time window, the statistics of various early warning logs are used to describe the operating characteristics of the system, and use this as a decision feature for failure warning. Correspondingly, two sets of evaluation systems are used. One set is the error cost function commonly used in traditional time series analysis methods, including mean square error (MSE), mean absolute error (MAE), and mean absolute ratio error (MAPE). Another set of evaluation systems uses the predicted accuracy, recall (Recall) and F-value (F-Measure) as evaluation indicators.

4. Conclusioon

Under the background of the rapid development of big data technology, in the face of increasingly complex network system operation and maintenance, using big data technology to carry out log analysis is a development trend and an indispensable and important means. The massive log real-time processing system based on Spark Streaming designed in this paper has realized the core function of the whole process from original log collection to calculation, analysis and storage. Meanwhile, the system has good stability and reliability, which can effectively improve the efficiency of operation and maintenance personnel, and is of great significance for operation and maintenance management in the age of big data.

References

[1] Pavel Smirnov,Mikhail Melnik,Denis Nasonov. Performance-aware scheduling of streaming applications using genetic algorithm[J]. Procedia Computer Science,2017,108:.

[2]Lekha R. Nair,Sujala D. Shetty,Siddhanth D. Shetty. Applying spark based machine learning model on streaming big data for health status prediction[J]. Computers and Electrical Engineering,2017.

[3]Liang Y, Zhang Y, Xiong H, et al. Failure Prediction in IBM Blue Gene/L Event Logs[C] //International Conference on Data Mining. Washington D.C.,USA: IEEE Computer Society, 2007: 583-588.

[4]Gujrati P, Li Y, Lan Z, et al. A Meta-Learning Failure Predictor for Blue Gene/L Systems[C]//International Conference on Parallel Processing. Washington D.C.,USA: IEEE Computer Society, 2007: 40-40.

[5]Lan Z, Gu J, Zheng Z, et al. A study of dynamic meta-learning for failure prediction in large-scale systems[J]. Journal of Parallel and Distributed Computing, 2010, 70(6):630-643.

[6]Fronza I, Sillitti A, Succi G, et al. Failure prediction based on log files using Random Indexing and Support Vector Machines[J]. Journal of Systems and Software, 2013, 86(1): 2-11.

[7]Gurumdimma N, Jhumka A, Liakata M, et al. Towards Detecting Patterns in Failure Logs of Large-Scale Distributed Systems[C]//Parallel and Distributed Processing Symposium Workshop. Washington D.C.,USA: IEEE Computer Society, 2015: 1052-1061.

[8]Nakka N, Agrawal A, Choudhary A. Predicting Node Failure in High Performance Computing Systems from Failure and Usage Logs[C]//IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum. Washington D.C.,USA: IEEE Computer Society, 2011: 1557-1566.

[9]Yu L,Zheng Z, Lan Z, et al. Practical online failure prediction for Blue Gene/P: Period-based vs event-driven[C]// The International Conference on Dependable Systems and Networks Workshops. Washington D.C.,USA: IEEE Computer Society, 2011:259-264. [23]Fu X, Ren R, Zhan J, et al Log Master: Mining Event Correlations in Logs of Large-Scale Cluster Systems[C]//IEEE International Symposium on Reliable Distributed Systems. Washington D.C.,USA: IEEE Computer Society, 2012: 71-80.

[10]Tang L, Li T, Perng C S. Log Sig:Generating System Events from Raw Textual Logs[C]//ACM International Conference on Information and Knowledge Management. New York,USA: ACM Press, 2011: 785-794.

[11]Fu Q, Lou J G, Wang Y, et al. Execution Anomaly Detection in Distributed Systems through Unstructured Log Analysis[C]//IEEE International Conference on Data Mining. Washington D.C.,USA: IEEE Computer Society, 2009: 149-158.