Research on the Auxiliary Application of Index Correlation Algorithm Analysis in the Fault Tracing of Finished Product Tobacco Sorting

Bailin Pan, Huan Le, and Yumin Wang

Hangzhou Cigarette Factory, China Tobacco Zhejiang Industrial Co., LTD,

Hangzhou, 310024, China

Keywords: Time Series Data, Association Algorithm, Anomaly Location, Semi Supervised Machine Learning, Fault Backtracking

Abstract: Taking Hangzhou Cigarette Factory of Zhejiang Zhongtobacco Industry Co., Ltd. as an example, Aiming at the back flow fault of finished tobacco scanning and sorting in the cigarette factory's packaging logistics production line, through the construction of the auxiliary system for the back flow fault tracing of finished tobacco scanning and sorting, the algorithm analysis of the relevant indicators of the production environment is carried out, the alarm convergence ability is improved, the root cause of the fault is quickly located, the steady-state model of the production business is established, and the fault tracing function is realized. Improve the ability of production data visualization, production process transparency and production decision-making intelligence. The back flow fault of finished tobacco code scanning and sorting is caused by the jitter of environmental factors caused by PLC transmission mechanism. In this study, various monitoring indicators related to production environment are collected through various technical tools, such as ZABBIX, APM, network packet capturing, application buried point monitoring, etc., which are usually stored as time series data (including collection time and indicator value), and a large number of indicators will be collected The historical data and real-time data of are imported to the big data platform to clean, store, analyze and mine the environmental data related to the production process of finished tobacco scanning and sorting. At the same time, an end-to-end model that can reflect the relationship between topology strength and weakness is constructed through the application performance database of traceable volume package logistics combined with the business architecture for visual display. It provides ideas for the intelligent construction of operation and maintenance of production system in cigarette industry.

1. Introduction

This paper takes Hangzhou cigarette factory of Zhejiang China Tobacco Industry Co., Ltd. as an example, aiming at the return fault of finished product tobacco sorting and sorting in the cigarette factory's packaging logistics production line, through the construction of the auxiliary system for the return fault tracing of finished product tobacco sorting and sorting, the algorithm analysis of the relevant indicators of the production environment is carried out to improve the alarm convergence ability, quickly locate the root cause of the fault, and establish the stable production business Model to realize the function of fault tracing. Improve the ability of production data visualization, production process transparency and production decision-making intelligence.

The back flow fault of finished tobacco code scanning and sorting is caused by the jitter of environmental factors caused by PLC transmission mechanism. In this study, various monitoring indicators related to production environment are collected through various technical tools, such as ZABBIX, APM, network packet capturing, application buried point monitoring, etc., which are usually stored as time series data (including collection time and indicator value), and a large number of indicators will be collected The historical data and real-time data of are imported to the big data platform to clean, store, analyze and mine the environmental data related to the production process of finished tobacco scanning and sorting. At the same time, an end-to-end model that can reflect the relationship between topology strength and weakness is constructed through the application

DOI: 10.25236/mmmce.2020.012

performance database of traceable volume package logistics combined with the business architecture for visual display.

2. Background and Requirements of the Project

With the upsurge of "industry 4.0" and "industrial Internet", China's manufacturing industry is in a critical period of adjustment and transformation of traditional industries. In February 2018, the National Tobacco Monopoly Bureau released the integration of cigarette manufacturing and Internet in the tobacco industry, comprehensively promoting the deep integration of new technologies such as cloud computing, big data, artificial intelligence, industrial Internet and the tobacco manufacturing industry, clearly proposing the innovative application of new technologies in the manufacturing field, forming a new pattern of collaborative development of platform, data, application, service and security, driving and leading High quality development of tobacco intelligent manufacturing. Although tobacco manufacturing enterprises have achieved a high degree of automation, there is still a lot of room for improvement in the overall state perception of the production process. How to realize the cloud data and the overall digitalization of the whole process is a big problem faced by tobacco enterprises. We need to make full use of new technologies such as big data analysis and AI intelligent decision-making to improve business and management capabilities.

As a common fault in the production line of coil and Parcel Logistics, the back flow fault of finished tobacco scanning and sorting is caused by the jitter of environmental factors caused by the transmission mechanism of production PLC. Due to the complex production environment factors, the specific causes of the fault are also different, such as the performance bottleneck of firewall, heartbeat timeout of database cluster, IO delay of storage disk, etc. When the fault occurs, there will be the phenomenon of scanning, sorting and backflow of finished cigarettes, and a large number of finished cigarettes will jump out of the production line, resulting in economic losses. In this study, starting from the back flow fault of finished product cigarette scanning, sorting and sorting, the intelligent transformation of cigarette production system in operation and maintenance management direction has been completed through KPI data collection, KPI data storage and operation, research and application of related algorithms and machine learning, and the construction of fault tracing system, which has taken an important step for the integration of cigarette manufacturing and Internet in tobacco industry.

3. System Construction Route

The system construction route is shown in the following figure:

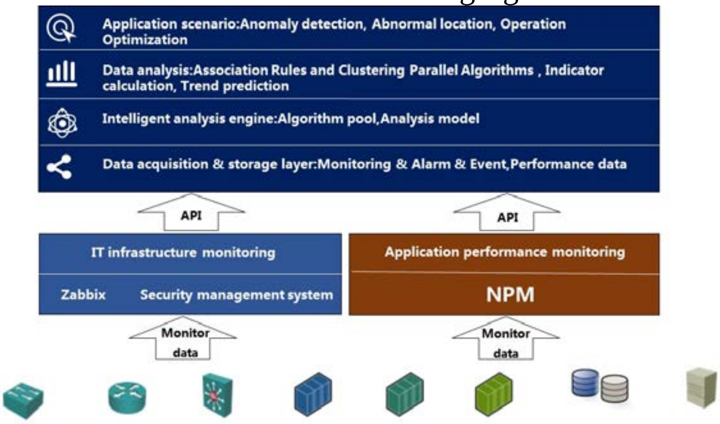
Application scenario:Anomaly detection, Abnormal location, Operation Optimization
Data analysis:Association Rules and Clustering Parallel Algorithms , Indicator calculation, Trend prediction
Intelligent analysis engine:Algorithm pool,Analysis model
Data acquisition & storage layer:Monitoring & Alarm & Event,Performance data
API
API
IT infrastructure monitoring
Zabbix
Security management system
Application performance monitoring
NPM
Monitor data
Monitor data

Fig.1 System constructionplan

3.1 Data Acquisition and Storage

Based on the requirements of data traceability and association algorithm, the data architecture is divided into data access, data bus, data calculation and data storage.

In the data bus layer, after the data enters the message bus (Kafka), ETL processing, filtering, segmentation, expansion and other operations of the data are carried out. The processed data enters the Kafka cluster again for streaming calculation. The real-time calculation adopts kafkastream, storm and other frameworks, and the off-line operation adopts spark.

In data computing layer, data computing mainly provides real-time computing and offline computing of data. Real time calculation, such as indicator exceptions, log keywords and other real-time alarm functions, offline data, such as data baseline, prediction, etc. MapReduce is used for common ETL related data processing tasks of heavy IO type, while MPI and spark are mainly used for algorithm model training.

The data storage layer stores different storage methods according to different application scenarios and data, stores them in ES for full-text search, performance data in TSDB, relationship data in mysql, complex relationship, historical data in HDFS, and provides offline calculation data.

3.2 Research on Index Relevance Algorithm

In the research of index relevance algorithm, the paper search and historical data import test are mainly used to select the suitable algorithm in the current time series algorithm which is commonly used in index screening, and the relevant interface is developed in the big data platform for algorithm import.

The overall architecture design is as follows:

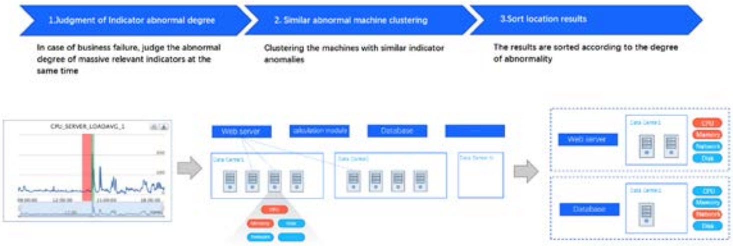
1.Judgment of Indicator abnormal degree
In case of business failure, judge the abnormal degree of massive relevant indicators at the same time
2. Similar abnormal machine clustering
Clustering the machines with similar indicator anomalies
3.Sort location results
The results are sorted according to the degree of abnormality
CPU_SERVER_LOADAVG_1
Web server
calculation module
Database
Data Center1
Data Center2
CPU
Memory
Network
Disk

Fig.2 The overall architecture design

In case of failure, the positioning system will be triggered to start analysis. It is mainly divided into three parts:

(1) Evaluation of indicator anomaly degree: the system will collect indicator data of all machines and modules in the current period of time, and execute anomaly detection algorithm to detect the anomaly degree of all indicators. The purpose of anomaly detection algorithm is to detect the anomaly pattern of index time series curve. In this project, the anomaly detection problem is transformed into a statistical probability observation model. For a single KPI, the initial anomaly degree score is calculated by kernel density estimation algorithm, and then the problem caused by periodic mutation is solved by extreme value theory, and the final anomaly degree score is given. Due to the periodic mutation of the data, the confirmation of the operation and maintenance personnel meets the expectation, and can not be judged as abnormal, such as the abnormal mutation caused by the early, late peak and real-time level start event. In order to solve this problem, the algorithm adds the method of extreme theory to make judgment. If it is a periodic mutation, it cannot be judged as an exception.

(2) Similar abnormal machine clustering: after getting the abnormal degree of all indexes, the machines with similar abnormal indexes are clustered by clustering algorithm. The purpose of similarity anomaly machine clustering algorithm is to integrate the machines with similar anomaly indexes and reduce the burden of troubleshooting. The core idea of the algorithm is to make the abnormal index information of each machine into a vector, and design a suitable clustering algorithm to do machine clustering. The clustering algorithm consists of three parts: input vector, distance function and clustering algorithm. This algorithm defines the input vector as (o 0, u 0, O 2, u 2 O - n, u - n), where o - N and u - n represent the abnormal degree of the nth KPI's sudden increase and decrease respectively. The distance function is used to calculate the clustering of two vectors. The closer an example is, the more similar the two vectors are. Finally, the Pearson

correlation algorithm is used in this project. After defining the vector and distance function, the clustering algorithm uses DBSCAN algorithm, which is a spatial clustering algorithm based on density. DBSCAN algorithm has the advantages of fast clustering speed and can effectively deal with noise points and find spatial clustering of arbitrary shape.

(3) Sorting of positioning results: by running the intelligent sorting algorithm, all clustering results are ranked according to the degree of abnormality, and finally presented to the operation and maintenance management personnel. Finally, the location system of this project will sort the indicators of different categories, using the current advanced search engine sorting technology learning to rank (learning to rank) in pointwise. By using the method of logical regression of semi supervised machine learning and learning the results of some manual annotation, a proper sorting model is automatically trained.

3.3 Intelligent Display

In terms of system presentation, an end-to-end model that can reflect the relationship between topology strength and weakness is constructed by using traceable volume package logistics application performance database and business architecture for intelligent visualization.

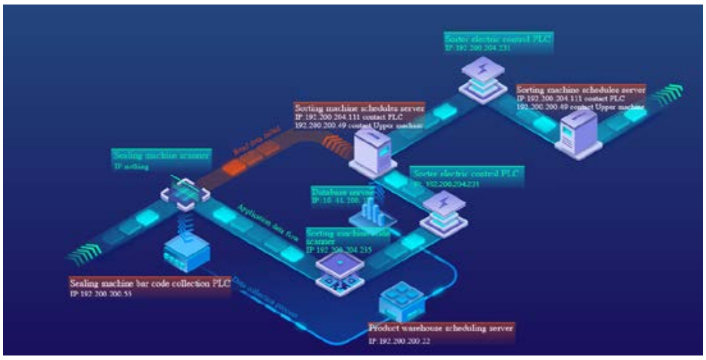
Sorter electric control PLC
IP:192.200.204.231
Sorting machine schedules server
IP:192.200.204.111 contact PLC
192.200.200.49 contact Upper machine
Sorting machine schedules server
IP:192.200.204.111 contact PLC
192.200.200.49 contact Upper machine
Sealing machine scanner
IP nothing
Sorter electric control PLC
IP: 192.200.204.231
Database server
Application data flow
Sorting machine scanner
IP:192.200.204.235
Sealing machine bar code collection PLC
IP:192.200.200.55
Product warehouse scheduling server
IP:192.200.200.22

Fig.3 Finished product tobacco scanning and sorting production model

The data analysis algorithm platform will display and cut the anomaly detection, root cause positioning and other results into independent components, which can be relatively freely embedded in the display platform. In the process of configuring the rest API and streaming, an additional Kafka message queue is specified. When an exception event is generated, the rest API / flow algorithm analysis process will push the brief content and index of the exception event to the specified message queue. After the alarm display platform subscribes to the message index pushed by Kafka, if necessary, the alarm details can be obtained through the rest API interface. In the rest API and flow processing flow, the system has reserved a hook entry, which can mount programs to call the alarm monitoring platform or other three-party platform APIs, and output the alarm and analysis detailed data to the display system or do some customized development work.

4. Application Profit

(1) Based on the detailed and summarized data, the intelligent algorithm is added for multidimensional analysis and data mining, which creates favorable conditions for operation and maintenance innovation;

(2) Through comprehensive data collection and data centralization, it provides a consistent data base for management analysis and mining fault prediction. Through the training of unsupervised and semi supervised machine learning on algorithm model, it provides the operation and maintenance personnel with the auxiliary function of intelligent decision-making and analysis, quickly locates the fault location, greatly shortens the time of troubleshooting, and improves the efficiency of operation and maintenance;

(3) It improves the recovery target time of the system and reduces the probability of production business interruption;

(4) In the medium and long term, the construction of the system data platform can integrate and

clean the data scattered in various business systems, improve the overall data quality and give full play to the business value of the data.

References

[1] New generation AI development plan (GF [2017] No. 35), State Council, http://www.gov.cn/zhengce/content/2017-07/20/content_.htm

[2] P. Bahl, R. Chandra, A. Greenberg, S. Kandula, D. A.Maltz, and M. Zhang. Towards highly reliable enter-prise network services via inference of multi-level dependencies. In SIGCOMM, 2007.

[3] M. Basseville, I. V. Nikiforov, et al. Detection of abrupt changes: theory and application, volume 104.Prentice Hall Englewood Clis, 1993.

[4] D. J. Berndt and J. Cliord. Using dynamic time warping to find patterns in time series. In Knowledge Discovery and Data Mining, pages 359{370, 1994.

[5] Y. Chen, B. Hu, E. Keogh, and G. E. Batista. Dtw-d:time series semi-supervised learning from a single example. In KDD, pages 383{391. ACM, 2013.

[6] I. Cohen, J. S. Chase, M. Goldszmidt, T. Kelly, and J. Symons. Correlating instrumentation data to system states: A building block for automated diagnosis and control. In OSDI, pages 231{244, 2004.

[7] I. Cohen, S. Zhang, M. Goldszmidt, J. Symons,T. Kelly, , and A. Fox. Capturing, indexing, clustering, and retrieving system history. In Proc.SOSP, pages 105{118, 2005.

[8] J. Cohen. Statistical power analysis for the behavioral sciences. 1988.

[9] Q. Fu, J.-G. Lou, Q.-W. Lin, R. Ding, Z. Ye,D. Zhang, and T. Xie. Performance issue diagnosis for online service systems. In SRDS, October 2012.

[10] A. Gretton, K. M. Borgwardt, M. Rasch,B.Scholkopf, and A. J. Smola. A kernel method for the two-sample-problem. volume 19, page 513. MIT;1998, 2007.

[11] B. Gruschke et al. Integrated event management:Event correlation using dependency graphs. In Proc.DSOM 98, pages 130{141, 1998.