Operation Anomaly Monitoring of Customer Service Data Analysis Platform Based on Improved FP-Growth Algorithm

Jing Yang 1*, Lihua Gong 1, Kunpeng Liu 1 and Letian Xiu 1

1 State Grid Customer Service Center, Tianjin, China
*Corresponding author’s e-mail: jing-yang@sgcc.com.cn

Abstract. Aiming at the problems of long time-consuming monitoring and poor monitoring accuracy in traditional customer service data analysis platform operation abnormality monitoring methods, a customer service data analysis platform operation abnormality monitoring method based on the improved FP-Growth algorithm is designed. Obtain customer service data sets, classify data types, filter customer behavior, identify the operating status of the data analysis platform, improve the FP-Growth algorithm to build a rule configuration model, set the platform safety factor threshold, and keep the reconstruction error of customer service data to a minimum. Within the scope, optimize the abnormal monitoring mode. The experimental results show that the average recovery time of the proposed customer service data analysis platform operation anomaly monitoring method is 5.239s, and the average platform operation anomaly accuracy rate is 97.3%, indicating that the customer service data analysis platform integrated with the improved FP-Growth algorithm operates abnormally. The monitoring method performs better.

1. Introduction
The improved FP-Growth algorithm is an algorithm used to describe the mining of frequent patterns without any candidate set [1,2]. According to the principle of data set classification and processing, after data scanning is completed, the extracted frequent sets are compressed until a frequent pattern tree with more branches is formed. Then, mining is carried out one by one according to the associated information between the frequent sets of data. The purpose of applying the improved FP-Growth algorithm to the abnormal monitoring of customer service data analysis platform operations is to explore a more complete data processing model. In addition to the application of platform operation abnormality monitoring in the field of fault diagnosis, it can also be used in fields such as mail filtering. It is currently one of the more widely used technologies. The customer service data analysis platform is based on the corresponding data processing technology. At the same time, when the amount of data shows a large-scale and rapid growth trend, the various types of frameworks and components of the customer service data analysis platform open source processing Tasks also need to be constantly adjusted and upgraded. However, during operation, in addition to the explosive growth of data, there are also potential risks including network attacks, server paralysis, and hardware crashes. Therefore, it is very necessary to monitor the abnormal operation of the customer service data analysis platform. The security of the current customer service data analysis platform is mainly provided by some infrastructure, and the upper-layer mechanism of the platform is relatively safe. However, for the lower-level facilities of the platform, the coverage of network abnormality monitoring cannot meet the security requirements. At this time, the platform architecture needs to be decomposed and optimized. Due to the particularity of the customer service data analysis platform itself, the technology for...
abnormal monitoring from the perspective of the server host log is relatively mature, but for the continuity and variability of data, it cannot complete data processing at one time. It is necessary to incorporate more advanced technologies at the data classification level. Academia has also accumulated some research results on abnormal monitoring, mainly focusing on the safe processing of data. In reference to the problem of temporal data congestion in the Industrial Internet of Things, the literature [3] disassembled the time series into multiple segments, and obtained a standardized process for data analysis; based on the calculation results of the correlation matrix of the data analysis, the time series correlation was constructed Graph model, and experimentally verify the time series in the model to obtain monitoring results, but in this process, the problem of unclear operation status is ignored. For the network security threat problem, literature [4] adopts the method of data segmentation pattern matching and semantic analysis to optimize and upgrade, extract the attack type, and locate the malicious code existing in the network level. After experimental testing, it is obtained. The experimental results are presented, but the description of the problem of unclear operation status still lacks more dimensions. Therefore, this paper proposes a method for monitoring abnormal operation of customer service data analysis platform based on the improved FP-Growth algorithm, and perfects the related topics of abnormal monitoring of customer service data analysis platform operation.

2. Operation anomaly monitoring of customer service data analysis Platform based on improved FP-Growth algorithm

2.1. Getting the customer service data set
Customer service data is mainly a platform-centered data set, which serves as the operation carrier of the whole platform [5,6]. Oriented by customer service needs, the corresponding statistical mode of data and statements is formed, which mainly includes topics such as user complaints, business handling, outbound statistics and customer service, and also includes several statistical applications under the four topics. Combined with the overall call volume of customer service, employee output, and business inquiry and processing data, summarized and analyzed the corresponding customer service data characteristics. In the current era of constantly changing market environment, customer service data also needs to meet the needs of the development of The Times, constantly adjust and upgrade, to achieve data management support. Through the abstract management of customer service data, the abstract representation of data set is obtained. In addition to reflecting the customer service data itself, the existing customer service data can be summarized by means of computing [7,8,9]. And according to the data and information relationship between customer service data, divide the data type, get more accurate data structure. The paradigm expression between customer service data is a kind of characteristic data that directly reflects the data relationship. The main idea of the standardized processing of customer service data is to eliminate the non-conforming data in the data set according to the corresponding processing rules in order to ensure that the relationship between data can be kept within a certain distance. In the platform, the attribute values of the data set are not unique in terms of the use of customer service data and have different meanings in different scenarios. At the same time, the attribute types of customer service data are also divided into major attribute and non-major attribute. As for the dependency relationship between data, any independent non-primary attribute data is completely subordinate to the primary key structure of the whole platform, but it also retains certain non-primary attributes. Taking into account the overall scalability requirements of customer data, the corresponding investment costs and management factors are taken into account. At the data management mode level of the platform, the delivery dependency between data is truncated. Combined with the representative data conversion method, using the tool software with integration function, based on the concept of database, establish a data system in line with the operation concept of the platform. Based on the service demand of customer service data, rationalize the internal scheduling of the platform. The data acquisition channel is used to improve the application and processing mode of customer service data. For customer service data analysis platform, standard and
unified channel information, according to data application requirements and information retrieval requirements, the best solution.

2.2. Identifying operation status of data analysis platform

Generally, in some platforms with large data scale, the measurement method of the relationship between data will affect the overall operation status of the platform to a certain extent. In customer service data analysis platform, there is a linear relationship between some data sets and schedulable resources. Some external characteristics of customer service data, including performance indicators and availability indicators, can be used as a standard to measure the operating status of the platform. Given the combination of the number of platform components and space footprint, choosing to monitor the health of all customer service data can have a speed impact. In order to avoid the situation of monitoring index redundancy, the correlation between data is combined with centralized processing.

When the platform operation is under normal operation conditions, the safety factor of the platform operation is set as 1, and the relationship between any two adjacent variables is positive correlation. In this case, there is a certain linear relationship between customer service data. When the operation status of the platform is abnormal, the value of the safety factor of the platform operation will change from 1 to -1, and the relationship between any two adjacent variables in the platform is completely negative correlation. According to this judgment standard, when the variable relationship is closely correlated, it indicates that the operation status of the platform is relatively normal, while when the variable relationship is alienated, it is necessary to carry out detailed comparison of various data of the platform operation. In order to reflect the impact of customers’ recent behaviors on the operating status of the platform, the weight of customers' behaviors is set after filtering customers' behaviors based on their behavior characteristics. The specific expression formula is as follows:

\[
L = \frac{1}{1 + \delta |r_m - r_n|} \tag{1}
\]

In formula (1), \(\delta\) represents the sequence of customer service sessions, \(r\) represents the attenuation coefficient of the platform, \(m\) represents the early data of customer service, and \(n\) represents the current data of customer service. According to the attributes of sample variables in the data, the orthogonal matrix expression formula of platform operation data is obtained:

\[
D = \sum d^{-1}
\tag{2}
\]

In formula (2), \(d\) represents the new random variable and \(z\) represents the dimension. The operating state of the platform is not constant across different uptime and application scenarios. In different state changes, Nengou captures some laws of customer service data analysis platform. The improved FP-growth algorithm is used to mine the historical operational metrics of the platform and use the aggregated data as the underlying data set. Based on the historical data of platform operation, reliable conclusions are drawn. The threshold setting of safety factor is also very important in the process of platform operation status identification. If the threshold range is higher than the actual value, the coefficients that should be relevant are more likely to be misjudged as irrelevant. Therefore, it is necessary to judge whether human intervention is needed according to the real-time operation of customer service data analysis platform. Based on this, the steps of identifying the operation status of the data analysis platform are completed.

2.3. Improve FP-growth algorithm to build rule configuration model

The improved FP-growth algorithm is an algorithm that upgrades the search strategy on the basis of the FP-growth algorithm and has the advantage of divide and conquer according to different types of data [10,11]. In the process of mining frequent itemsets, scan data, customer service will be a frequent
itemsets, and then according to the setting of threshold platform security coefficient, will not satisfy the operation of the platform delete data sets, then get the new data set, and according to the new standard data set of frequent items, support count of descending order. In the actual application process, in order to improve the defect of scanning customer service database times, the improved FP-growth algorithm is used to adjust the arrangement of customer service data [12,13,14]. At the same time, according to the algorithm only need to traverse the characteristics of the database, separate from the original database for secondary mining, do not need to mining platform database again. In the process of building the rule configuration model, it is necessary to implement the rule from top to bottom in the order of conceptual model design, logical model design and physical model design step by step. Random vectors in customer service data are collected as the matrix of data samples, and the specific expression formula is as follows:

\[
Q = \begin{bmatrix}
q_{11} & q_{12} & \cdots & q_{1p} \\
q_{21} & q_{22} & \cdots & q_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
q_{l1} & q_{l2} & \cdots & q_{lp}
\end{bmatrix}
\]  

(3)

In formula (3), \(q\) represents the standard deviation of the random variable, \(p\) represents the coefficient vector, and \(l\) represents the load. On the basis of formula (3), model elements are transformed and expressed as follows:

\[
k_{ij} = \begin{cases}
x_{ij}, & 0 \\
-x_{ij}, & 1
\end{cases}
\]  

(4)

In formula (4), \(x\) represents the opposite indicator, \(-x\) represents the opposite indicator, and \(ij\) represents the weight of the indicator. Based on the calculation results, you can determine the service coverage of the rule configuration after analyzing the platform operation requirements. On the basis of high generalization, the customer service data objects involved in the relevant business are classified. After dividing the subject domain, the subject of the data set is refined, and the rule attributes corresponding to each data set are clarified. Based on this, the steps of improving FP-growth algorithm to build the rule configuration model are completed.

2.4. Optimizing the anomaly monitoring mode

The anomaly monitoring mode of the customer service data analysis platform can discover the potential risks of the platform operation in a short period of time and take effective measures in time. In the process of platform operation, once the actual situation does not match the expected behavior, the problems are defined and the abnormal situation is described [15]. If an abnormal data is prominent in the sample set to be monitored, its observed value will also exceed the standard threshold range. Using the principle of the improved FP-growth algorithm, suspicious data can be found and marked for early warning. If the current statistical data are far from the expected outlier, then relevant effective data are defined, and the expression formula for the definition of mean value of abnormal data is as follows:

\[
\lambda = \sum_{c=1}^{e} \frac{\hat{h}_c}{c}
\]  

(5)

In formula (5), \(\hat{h}\) represents the data set to be monitored, \(c\) represents the expected value of data outlier, and \(e\) represents the outlier. The purpose of defining the mean value of abnormal data is to simplify the original data of the platform and keep the reconstruction error of customer service data within the minimum range. In addition to the orthonormal vectors between data, noise and redundancy between data are the factors that affect the accuracy of monitoring. In the heterogeneous space of
platform anomaly monitoring, the larger the volume of corresponding diagonal elements is, the greater
the noise of customer service data is in the layer network, and the commonality of data redundancy is
relatively high to a certain extent. Conversely, the smaller the element, the less likely it is to prove that
there is noise and data redundancy. At the same time, a group of obvious linear feature vectors is found
through linear transformation, and the extracted classification information is compressed to express
the dimension of feature space in the form of the largest category distance. On the basis of high
probability distribution, the expression formula of criterion function is obtained:

\[ G(\eta) = \frac{\eta^R W_s}{\eta^S} \quad (6) \]

In formula (6), \( \eta \) represents the total number of samples, \( W \) represents the maximum inter-class
hash degree, \( s \) represents the signal separation speed of the data analysis platform, and \( R \)
represents the minimum intra-class dispersion. According to the calculation results of Formula (6), the
expression formula of projection coefficient of independent basis vector is obtained:

\[ \mathcal{E} = \frac{F_{T-1}}{V_T} \quad (7) \]

In formula (7), \( F \) represents the linear combination coefficient, \( V \) represents the number of test
data, and \( T \) represents the separated signal source of the data analysis platform. According to the
calculation results of formula (5) to formula (7), anomaly monitoring steps in line with the platform
operation and development law are formulated for data acquisition, modeling and monitoring stages
respectively. Combined with the fault index and abnormal probability of anomaly monitoring, the
corresponding parameter updating mechanism is integrated. In addition to detailed monitoring of the
platform's internal operations, the operating environment outside the platform should also be
monitored to ensure that the customer service operation platform is free from external interference.
Based on the above description, complete the steps of optimizing the anomaly monitoring mode.

3. Experimental test

3.1. Building experimental environment

In order to verify the effectiveness of the operation anomaly monitoring method of the customer
service data analysis platform designed this time, the experimental test was carried out and the
corresponding experimental environment was built. Operating platform Select Linux platform, and
take Map operator and Join operator as data set of operation anomaly monitoring. The experimental
computing framework chooses Apache Flink as the computing engine with good performance and
uses Nmon as the open source data tool. Nmon can automatically integrate the information of the
platform server and the usage of data resources, and summarize them into a specific file according to a
fixed format. It is a common analysis and use tool for monitoring platform operation anomalies in
experimental environments. In addition, during the experiment, Nmon system has a low occupancy
rate and does not cause system congestion. Under normal operating conditions, Nmon's CPU usage is
between 1% and 2%. The software and hardware configuration of the experimental environment is
shown in Table 1:

| Serial number | Hardware | configure | Software |
|---------------|----------|-----------|----------|
| #1            | disk     | 600G, 7200RPM | Flink    |
| #2            | network  | 100MB/s    | Zookeeper|
| #3            | Memory   | 16GB       | MySql    |
| #4            | OS       | Ubuntu     | Nmon     |
| #5            | CPU      | Intel(R)Core(TM) 64.bit i5-6200 3GHz | IntelliJ IDEA |
On the basis of Table 1, select the data set required for this experiment, mainly the real index data collected through nmon. Among them, there are 22300 normal data, accounting for 82% of the data set, and 7433 abnormal data, accounting for 24% of the data set. In addition, multiple indicators are set for the processor, disk, network memory and other resources of the platform, mainly including CPU, disk, I/O and memory. The relevant indicators and descriptions are shown in Table 2:

| Indicator type | Indicator name | Index description |
|----------------|----------------|-------------------|
| CPU            | QL-memtotal    | Physical memory occupies CPU space |
|                | QL-memfree     | Free memory occupies the total CPU space |
|                | QL-membuff     | The memory space of the kernel cache |
|                | QL-memcach     | Total cache memory occupies CPU space |
|                | QL-memutil     | Total physical memory space used by the platform |
|                | QL-swpin       | The amount of virtual memory space read from disk per second |
|                | QL-swptotal    | Total space of the platform exchange area |
| Disk and I/O   | ci-tps         | The number of requests sent per second by a data port |
|                | ci-svctm       | Port service time |
|                | ci-quesize     | Port queue Length |
|                | ci-read        | The number of bytes read per second by a platform port |
|                | ci-util        | An overview of platform operations |
|                | ci-reqsize     | Data volume size |
|                | qr-user        | User space Space occupied by users |
|                | qr-sys         | Kernel space takes up space |
|                | qr-hirq        | Hardware interrupt request time takes up space |
|                | qr-sirq        | Software interrupt request time occupies space |
| network        | Pz-packets     | The number of packets sent by the platform |
|                | tr-bytes       | The number of bytes sent by the output port |
|                | pz-dropped     | Sending Port Number of packets lost when sending data |
|                | tp-packets     | The number of packets received by the platform |

Combined with the data set information in Table 2, the monitoring indicators required by the experiment were measured, and experimental tests were conducted on the basis of reducing platform overhead.

3.2. Experimental result
The experimental scenario is as follows: the platform is injected with a single fault type, the corresponding monitoring period is set, and the CPU utilization in the monitoring period is collected. The collection time is set from 20:15:05 to 23:15:05, and the fault injection time is 15 minutes and 10
minutes respectively. The process deadlock is selected as the fault exception injection method for the three operation anomaly monitoring methods of customer service data analysis platform. During the experiment, the input rate of the tuple was adjusted to 500-1500 tuples/s, and the work task of the platform branch node was terminated. The operation anomaly monitoring method of customer service data analysis platform based on dynamic rule base and association rule are selected and compared with the operation anomaly monitoring method of customer service data analysis platform in this paper. Test the recovery time of three abnormal operation monitoring methods of customer service data analysis platform under different input rates. The experimental results are shown in Figure 1-3:

![Figure 1](image1.png)  
**Figure 1** Input rate 500 tuples/s recovery time (s)

![Figure 2](image2.png)  
**Figure 2** Input rate 1000 tuples/s recovery time (s)
Abnormal operation monitoring method of customer service data analysis platform based on dynamic rule base
Abnormal operation monitoring method of customer service data analysis platform based on association rules
The abnormal operation monitoring method of customer service data analysis platform in this paper

According to Figure 1-3, it can be concluded that when the input rate becomes 500~1000tuples/s, the recovery time difference of the three customer service data analysis platform operation abnormal monitoring methods is relatively small; when the input rate becomes 1500tuples/s, the recovery time gap of the three kinds of customer service data analysis platform operation anomaly monitoring methods is gradually increasing. From this, the average recovery time of the designed customer service data analysis platform operation abnormality monitoring method and the other two customer service data analysis platform operation abnormality monitoring methods are obtained, as shown in Table 3:

| Input rate(tuples/s) | The designed customer service data analysis platform operation abnormal monitoring method | Operation anomaly monitoring method of customer service data analysis platform based on dynamic rule base | Operation anomaly Monitoring method of customer service data analysis platform based on association rules |
|---------------------|------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------|
| 100                 | 1.321                                                                              | 1.983                                                                                          | 2.036                                                                                          |
| 200                 | 1.416                                                                              | 2.372                                                                                          | 2.408                                                                                          |
| 500                 | 2.561                                                                              | 3.316                                                                                          | 3.322                                                                                          |
| 700                 | 4.321                                                                              | 5.069                                                                                          | 5.834                                                                                          |
| 900                 | 5.644                                                                              | 6.358                                                                                          | 6.889                                                                                          |
| 1000                | 7.403                                                                              | 8.032                                                                                          | 8.512                                                                                          |
| 1200                | 8.226                                                                              | 9.784                                                                                          | 10.205                                                                                         |
| 1500                | 11.021                                                                             | 12.204                                                                                         | 11.819                                                                                         |

It can be seen from Table 3 that the average recovery time of the designed customer service data analysis platform operation abnormal monitoring method and the other two customer service data analysis platform operation abnormal monitoring methods are: 5.239s, 6.140s, 6.378s, indicating the design The customer service data analysis platform operates an abnormal monitoring method, and it takes less time to recover from an abnormal state.
The accuracy of the abnormal monitoring of the operation of the three customer service data analysis platforms was tested under the above test environment and conditions, and the results of the abnormal monitoring of the operation of the customer service data analysis platform were obtained as shown in Table 4.

Table 4 Anomaly monitoring accuracy rates of three customer service data analysis platform operations anomaly monitoring methods

| Number of experiments | The designed customer service data analysis platform operation abnormal monitoring method | Operation anomaly monitoring method of customer service data analysis platform based on dynamic rule base | Operation anomaly monitoring method of customer service data analysis platform based on association rules |
|-----------------------|---------------------------------|-------------------------------------------------|-------------------------------------------------|
| 1                     | 97.5%                           | 83.6%                                           | 79.6%                                           |
| 2                     | 97.6%                           | 82.5%                                           | 78.5%                                           |
| 3                     | 97.4%                           | 81.7%                                           | 77.9%                                           |
| 4                     | 96.5%                           | 83.5%                                           | 78.0%                                           |
| 5                     | 96.8%                           | 84.6%                                           | 80.5%                                           |
| 6                     | 97.2%                           | 80.5%                                           | 79.1%                                           |
| 7                     | 97.8%                           | 84.1%                                           | 78.6%                                           |
| 8                     | 97.5%                           | 85.2%                                           | 77.8%                                           |

It can be seen from Table 4 that the designed customer service data analysis platform operation abnormality monitoring method and the other two customer service data analysis platform operation abnormality monitoring methods have average abnormal monitoring accuracy rates of 97.3%, 83.2%, 78.6%, respectively. The designed customer service data analysis platform operation abnormality monitoring method has better abnormality monitoring accuracy.

4. Conclusion

The customer service data analysis platform designed in this paper is built on the basis of customer service data, using the improved FP-Growth algorithm to carry out the work of frequent itemsets mining in more detail. Experimental tests have proved that this method has improved the traditional platform operation abnormal monitoring method that takes too long to recover from abnormal conditions, and the monitoring accuracy of abnormal operation conditions is higher, which lays a foundation for the academic community to carry out research on related topics. The theoretical basis and practical basis of the platform, and promote the development process of the field of platform operations. Due to the limited research conditions, the article has not divided the scenes of abnormal monitoring of platform operations in detail. The relevant details will be continuously improved in the future.

References

[1] Ls A,  Fv C,  Rya B,  et al. (2020) A guided FP-Growth algorithm for mining multitude-targeted item-sets and class association rules in imbalanced data[J]. Information Sciences, doi: 10.1016/j.ins.2020.10.020

[2] Wu Y,  Zhang J. (2020) Building the electronic evidence analysis model based on association rule mining and FP-growth algorithm[J]. Soft Computing, 24(2), doi : 10.1007/s00500-019-04032-0

[3] Ding Xiaoou, Yu Shengjian, Wang Muxian, et al. (2020) Anomaly Detection on Industrial Time Series Based on Correlation Analysis[J]. Journal of Software, 31(03): 726-747.

[4] Li Yuan, Wang Guoyin, Li Zhixing, et al. (2019) A Sequential Attention Based Convolutional Neural Network for Anomaly Detection[J]. Journal of Zhengzhou University(Natural
Science Edition), 51(2): 17-22.

[5] Misra A., Bhuiyan M., J. Mahmud, et al. (2019) Using Structured Representation and Data: A Hybrid Model for Negation and Sentiment in Customer Service Conversations[J]. Tenth Workshop on Computational Approaches to Subjectivity.

[6] Li Z., Guo H., Wang W M., et al. (2019) A Blockchain and AutoML Approach for Open and Automated Customer Service[J]. IEEE Transactions on Industrial Informatics, 2019:1-1.

[7] Black A. (2019) 'Bringing People Together' - How Microsoft Teams and Other Collaboration Tools are Transforming Customer Service[J]. Database and network journal, 49(3):7-7.

[8] Zhang C., Ren M. (2021) Customer service robot model based on e-commerce dual-channel channel supply coordination and compensation strategy in the perspective of big data[J]. International Journal of System Assurance Engineering and Management, 2021:1-11.

[9] Thiriku J M., Kegoro H O., Mandal P C. (2019) Strategic leadership on customer service delivery in the ministry of industry, investment and trade in kenya[J]. Noble International Journal of Business and Management Research, 2019, 3.

[10] Bagui S., Devulapalli K., Coffey J. (2020) A Heuristic Approach for Load Balancing the FP-Growth Algorithm On MapReduce[J]. Array, 2020:100035. doi: 10.1016/j.array.2020.100035

[11] Jia Y., Liu L., Chen H., et al. (2020) A Chinese unknown word recognition method for micro-blog short text based on improved FP-growth[J]. Pattern Analysis and Applications, 2020, 23(2). doi: 10.1007/s10044-019-00833-z

[12] Liu S., Yun S U., Zhang Y. (2019) Open-line Fault Diagnosis and Positioning Method for 10kV Power Distribution Network Branch Line Based on FP-growth Algorithm[J]. Power System Technology, 43(12), 4575-4892.

[13] Jiang X., Xiao X., Wen W., et al. Operation Optimization of Thermal Power Plant Based on Parallel FP-Growth Algorithm[J]. Power System Technology, 2019.

[14] Zhang J., Sun H., Sun Z., et al. (2020) Reliability Assessment of Wind Power Converter Considering SCADA Multistate Parameters Prediction Using FP-Growth, WPT, K-means and LSTM Network[J]. IEEE Access, 2020, PP(99):1-1, doi: 10.1109/access.2020.2992089.

[15] Liu Qinglian, Wang Xueping. (2019) Real Time Monitoring Simulation of Abnormal State of Large Data In Smart Grid[J]. Computer Simulation, 36(3): 364-367.