Method for Identifying Abnormal Access to Sensitive Data Based on Network Flow

Yong Ma¹, Jing Feng¹ and Jingyi Li²
Civil Aviation electronic technology co., ltd, 55 Qing liang West Road, Xin jin County, Chengdu City, Sichuan Province
Email: heimajushi@sina.com

Abstract. With the increasing scale and quantity of civil aviation information system, the data carried in the information system is becoming more and more important and sensitive. In the face of a large number of information systems, sorting out and monitoring whether passenger information in the information system has been accessed in violation of regulations is an urgent problem in the civil aviation industry. In this paper, the civil aviation business system is deeply studied, and a method of identifying abnormal access to sensitive data based on network traffic is proposed. Through the collection and analysis of network traffic, and then identify the sensitive information in the network traffic, and use k-means clustering method to classify the access behavior of sensitive information. Finally, combined with expert experience, the malicious access behavior for sensitive data is accurately identified.

1. Introduction
With the increasing scale and quantity of civil aviation information system, the data carried in the information system is becoming more and more important and sensitive. Passenger information as one of the core data of civil aviation, the amount of relevant data has been increasing explosively in recent years, but the network security incidents of passenger information leakage occur frequently. In order to protect the information security of civil aviation passengers and safeguard the legitimate rights and interests of passengers, the protection of passenger information by civil aviation units is under increasing pressure. In the face of a large number of information systems, sorting out and monitoring whether the passenger information in the information system is accessed illegally is one of the key issues of civil aviation units. In this paper, through the collection and analysis of network traffic, and then identify the sensitive information in the network traffic, and use k-means clustering method to classify the access behavior of sensitive information. Finally, combined with expert experience, the malicious access behavior for sensitive data is accurately identified. Due to the large amount of information system data, the application system format is not uniform, it is difficult to focus on the application system log audit analysis, the method proposed in this paper through the analysis and processing of network traffic, can achieve the centralized audit of the use of sensitive data of multiple application systems, greatly improve the ability of data security management of civil aviation.

2. Research Status of Sensitive Information Identification Based on Traffic
At present, the related research on traffic based sensitive information identification technology at home and abroad mainly focuses on DLP, and most of the research focuses on traffic encryption technology. For example, Mohammed ghouse [1] and others studied encryption in the process of data transmission to ensure data security Zhang [2] and other scholars’ research content is the application of data encryption technology in the process of network transmission to ensure the security of data.
In the aspect of sensitive data transmission monitoring, relevant institutions and organizations at home and abroad identify and track the use of sensitive data in network traffic and block the transmission of sensitive data by extracting feature strings. However, this method is based on fixed features and cannot dynamically identify and block illegal access behavior.

In terms of sensitive information identification, Guo Dongdan [3], an expert in the civil aviation industry, has provided an automatic sensitive data recognition model. The model uses pattern recognition, keyword matching + pattern recognition, NLP and other technologies to realize the identification of sensitive information in structured and non-institutional databases. The model has an important guiding role in the identification of sensitive data based on traffic in this paper.

With the rapid development of civil aviation business, the passenger information data involved in civil aviation information system is also growing rapidly. However, with the increasing requirements of privacy information protection at home and abroad, the protection of passenger information has become an important work in the process of business development of civil aviation units. Based on the above research results, this paper proposes a method to identify abnormal access of sensitive data based on network traffic.

3. Identification Method Model
In this paper, through the packet capture and analysis of the network data flow, and then use the sensitive data identification model to identify sensitive data, and use k-means clustering to classify the access behavior of sensitive data. Finally, combined with expert experience, the malicious access is accurately identified. The model of recognition method in this paper is as follows:

3.1. Data Capture
This paper uses the scapy library to capture the network data flow. Because the current business system of civil aviation mainly uses HTTP protocol, the HTTPS protocol can be transformed into
HTTP protocol traffic by deploying certificates, so only HTTP protocol traffic is captured in this paper.

Table 1. Data capture

| filterstr="http" |
|------------------|
| sniff(filter=filterstr, iface='eth0', count=0) |

3.2. Packet Parsing

There are two steps in the packet parsing phase:

- Analyze the content of the packet, identify whether the network packet contains sensitive data, and classify the sensitive data according to the method of sensitive data classification model. For civil aviation passenger information may include: name, ID card, home address, telephone number, wechat account number, email account number, bank card number, flight number, electronic ticket number, flight time, departure place, destination, etc.

- Implementation methods: (1) using pattern matching method to match the length, character type and format of data; (2) using keyword matching to identify the header of structured data; (3) using keyword + pattern matching to identify keywords in the file, such as QQ, wechat, mailbox and so on, and then pattern matching in the keyword attachment To identify whether sensitive information exists. (4) NLP recognition method is used to identify the name or address information in the packet.

For the packet containing sensitive data, other information of the packet is analyzed, including source address, destination address, access time, access account, token and so on.

Table 2. Packet parsing

| data = { |
| 'time': ip_layer.time, |
| 'ip': {'src': ip_layer.src, 'dst': ip_layer.dst} |
| 'tcp': {'seq': tcp_layer.seq, 'ack': tcp_layer.ack} |
| } |
| data['type'] = 'request' |
| http_header = pkt[HTTPRequest].fields |
| http_header.pop("Headers") |
| data['header'] = http_header |

The process of data packet and packet analysis is a continuous process. In this paper, we set up to collect 5 days of network traffic for later data analysis. In the test environment, a total of 15 G traffic is collected.

3.3. Automatic Classification Model

This part is the key content of this paper. This paper selects K-means clustering to classify the access behavior of sensitive data in the network. By analyzing the characteristics of sensitive data access time, number, frequency and IP address of sensitive data in the network, K-means clustering algorithm is used to automatically classify these data.

- The main advantages of K-means clustering algorithm are as follows:
  - The algorithm is fast and simple;
  - It has high efficiency and scalability for large data sets; The time complexity is nearly linear, and it is suitable for mining large-scale data sets. The time complexity of K-means clustering algorithm is O (NKT), where n represents the number of objects in the dataset, T represents the number of iterations, and K represents the number of clusters.

This part is described in detail in the following chapters.

3.4. Expert Verification

Expert verification is to manually verify the results of automatic classification of sensitive data access behavior in network traffic by using k-means clustering through expert experience, analyze and judge
the differences between different classifications, and finally find out the types of abnormal access and related feature information.

4. Access Behavior Classification Method Based on K-means Clustering Algorithm

In order to automatically identify the access behavior of sensitive data in the network, find abnormal network access behavior, and improve the ability of data security management. In this paper, K-means clustering algorithm is used to classify and identify the access behavior. This data source comes from the civil aviation related data, which has a certain industry representative.

4.1. Data Preprocessing

This data analysis only focuses on the source address, destination address, access time, access frequency, amount of data access, and access user's rights. The specific data preprocessing methods are as follows:

- **Data source address**: for this kind of data, due to the wide range of users to access, this time, only segment B data is used in data preprocessing, and then the data is processed in decimal system.
- **Data destination address**: this kind of address is mainly the server address, and the address is relatively concentrated. In this preprocessing process, it is mainly the data of the last two parts of IP address, and then decimal processing is performed.
- **Access frequency**: this visit frequency mainly refers to the number of times a user accesses sensitive data in 5 days. Each time the same user account logs in once and uses sensitive data to access behavior is recorded as one time.
- **Access data volume**: the number of entries each user accesses to sensitive data. In this paper, in order to simplify the calculation difficulty, as long as the sensitive data with the same characteristics in the accessed data is recorded as one item, and the data is not classified.
- **User access rights**: since different permissions are assigned to each user in the system, the user permissions are also included in the analysis scope in this data analysis.

Finally, the pre-processing data are unified in the order of magnitude, and the processed data are all in the range of four positive numbers.

4.2. K-means Clustering Algorithm

The basic principle of K-means clustering algorithm.

K-means clustering algorithm is to divide a large number of unlabeled datasets into several categories according to their inherent attribute characteristics, so that the internal data of each category has certain similarity. K-means clustering belongs to unsupervised learning method.

The K-means clustering model is as follows:

- Set the sample as \( S = x_1, x_2 \ldots x_m \). Then the calculation process of K-means clustering algorithm is as follows:
  - Determine the centers of K initial classes, \( \mu_1, \mu_2 \ldots \mu_k \).
  - For \( X_i \) of each sample, the nearest category is marked as:
    \[
    label_i = \arg\min_{1 \leq j \leq k} \| x_i - \mu_j \|
    \]
  - Then repeat the calculation until the change of category center is less than the set threshold.

4.3. Algorithm Implementation

This paper uses the scikit-learn module based on python, which is specially used for machine learning. Scikit-learn is a machine learning algorithm library implemented in Python. Sklearn can implement data preprocessing, classification, regression, dimension reduction, model selection and other common machine learning algorithms. Sklearn is based on numpy, SciPy, Matplotlib.

Numpy Python implementation of open source scientific computing package. It can define high-dimensional array objects, matrix calculation and random number generation functions.

Advanced scientific computing package implemented by SciPy python. It is closely related to numpy. SciPy generally manipulates numpy array for scientific calculation, so it can be said that it is
based on numpy. SciPy has many sub modules which can deal with different applications, such as interpolation, optimization algorithm, image processing, mathematical statistics, etc.

Matplotlib Python implementation of the drawing package. It is very simple to generate histogram, histogram, histogram and bar chart of power chart, and it can be used to generate bar chart, histogram and power chart.

| Table 3. Algorithm implementation |
|-----------------------------------|
| estimator = KMeans(n_clusters=5)  |
| cc=estimator.fit_predict(dataSet) |
| labelPred = estimator.labels_     |

The K-means clustering algorithm is easy to understand. In the implementation, kmeans library file in sklearn library is called to realize data clustering, and Matplotlib library is used to realize drawing.

4.4. Result

The K-means clustering method is used to automatically classify the sample set, which is set to 5 categories, 4 categories, 3 categories and 2 categories in the experimental process. The specific classification effect is shown in the figure below.

![Figure 2. Result](image)

According to the experimental results, 5, 4, 3 and 2 categories are analyzed, as shown in the figure. The experimental results show that the classification results have the following characteristics in the above four cases

- Most of them are relatively concentrated;
- The distance between different categories is generally far;
- The data of individual categories are mixed;
- Generally speaking, the consistency of classification results is very high.

Through in-depth analysis of the data, it is found that different types of access characteristics are more obvious
The access behavior from the Internet has obvious characteristics, and this kind of data is basically divided into one category;
- Intranet address access is relatively centralized;
- The characteristics of the amount of data accessed are obvious;
- The characteristics between low frequency and high frequency are obvious.
Based on these characteristics of data sources, the classification effect is highly consistent.

![Accuracy comparison](image)

**Figure 3.** Recognition effect

From the above figure, it can be found that when the category is 5, the overall recognition effect is better, keeping above 80%. When set to other values, the recognition effect has some fluctuations, and the overall recognition effect is not stable. The reason is that the abnormal access behavior is not obvious, which is submerged in the whole big data set.

Under this data, from the automatic classification effect, when the category is 5, it is easier to find various network access behaviors for sensitive data.

| source address | destination address | data volume | frequency | access rights |
|----------------|---------------------|-------------|-----------|---------------|
| 1008           | 2211.83             | 15833.33    | 91.83     | 1333.33       |
| 1006.19        | 1252.88             | 5172.84     | 832.35    | 1039.51       |
| 1002.28        | 2397.85             | 11705.92    | 265.39    | 1065.79       |
| 1002.45        | 1608.23             | 766.67      | 789.43    | 1000          |
| 1002.49        | 1608.22             | 7834.75     | 736.81    | 1136.75       |

From the chart, we can see that there is abnormal access behavior in the center of the following circle:

| source address | destination address | data volume | frequency | access rights |
|----------------|---------------------|-------------|-----------|---------------|
| 1008           | 2211.83             | 15833.33    | 91.83     | 1333.33       |

Table 4. K-means clustering results

Table 5. Abnormal access behavior 1

| source address | destination address | data volume | frequency | access rights |
|----------------|---------------------|-------------|-----------|---------------|
| 1008           | 2211.83             | 15833.33    | 91.83     | 1333.33       |

Table 6. Abnormal access behavior 1

| source address | destination address | data volume | frequency | access rights |
|----------------|---------------------|-------------|-----------|---------------|
| 1002.28        | 2397.85             | 11705.92    | 265.39    | 1065.79       |

Through the comparison and analysis of the data in Table 5 and table 5, combined with the application system, it is found that there is indeed illegal access behavior in the data subset with the center of the circle above, and the network traffic characteristics of illegal access are very obvious.
5. Conclusion
This paper puts forward a method to identify the abnormal access of sensitive data based on network traffic. This method can solve the problem that the civil aviation units cannot handle the passenger sensitive data access audit through the application log because of the large number of information systems. At the same time, through the monitoring and analysis of the passenger sensitive data in the network traffic, the abnormal access in the network can be found in time. Ask behavior to help civil aviation units improve their data security management capabilities. However, in the specific analysis, it is found that the characteristics of the data set used this time are obvious, which leads to the high consistency of multi categories when using k-means classification algorithm. In view of the above problems, through experiments and demonstration, this paper finds that the identification effect is better for small traffic or recorded files; however, in the case of large data flow, the above problems have a great impact on the identification results, which may lead to the phenomenon of missing reports, which needs further improvement and improvement.

6. References
[1] Ghouse M, Nene M J, Vembuselvi C. Data Leakage Prevention for Data in Transit using Artificial Intelligence and Encryption Techniques[C] 2019 International Conference on Advances in Computing, Communication and Control (ICAC3). 2019.
[2] Jing-Shu Z. The application and practice of Enterprise terminal Data leakage prevention[J]. Digital Technology and Application, 2018.
[3] Dongdan Guo, Hao Sun. An Automated Recognition Model for Sensitive Information. Journal of Physics: Conference Series, Volume 1575, 5th Annual International Conference on Information System and Artificial Intelligence 2020.
[4] Guo G N, Yun Y G, Chu M, et al. Application of Density-Based Adaptive K-Means Clustering Algorithm in Web Log Mining[J]. Advanced Materials Research, 2012, 433-440:5152-5156.
[5] Munther A, Razif R, Abualhaj M, et al. A Preliminary Performance Evaluation of K-means, KNN and EM Unsupervised Machine Learning Methods for Network Flow Classification[J]. International Journal of Electrical & Computer Engineering, 2015.
[6] Shi Y, Peng X, Li R, et al. Unsupervised Anomaly Detection for Network Flow Using Immune Network Based K-means Clustering[J]. 2017.
[7] Ying P, Jing-Zhang L, Hui-Juan L I. Clustering Analysis of Network Customer's Action Based on K-means Arithmetic[J]. Computing Technology and Automation, 2007.