Research and Visual Application of sensitive event monitoring technology based on Semantic analysis

Zhaowei Cui[a], Jie Cheng, Teng Lu, Ran Zhang, Jing Li
[a]18810609305@163.com

Information and Communication Branch of State Grid Corporation of China Beijing China

Abstract—With the development of enterprise informatization, under the influence of the epidemic situation in recent years, the mobile office mode has brought new challenges to the enterprise security work while improving the work efficiency. As a part of the enterprise's "big security", security is related to the survival and development of the company and the long-term stability of the country. Based on the status quo of security work, this paper analyzes many risks faced by security work, studies intelligent judgment technologies related to semantic analysis, realizes awareness and visual display of security sensitive events on the network side, and builds a network monitoring platform supporting security, which can accurately and efficiently serve the enterprise security management, greatly improving the technical support level of enterprise security work, Effectively prevent leakage of secrets.

Keywords: Text extraction; Semantic analysis; Intelligence; Visualization

1 INTRODUCTION

With the increasing integration of the world economy, information technology has been widely applied and developed in various industries, enterprises are undertaking more and more national tasks, and new media are rising rapidly. Online office and mobile office have gradually become a new working mode. In recent years, under the influence of the epidemic, the global economic situation is severe, and mobile office has become an inevitable choice for many enterprises to effectively carry out their work. However, while improving the efficiency of enterprises, the security work of enterprises has faced new challenges. The risk of disclosure of documents with important sensitive information is getting higher and higher [1]. In addition, the penetration and theft of secrets by domestic and foreign hostile forces is still serious, so it is urgent to strengthen the monitoring capacity of security sensitive events.

At present, the problem of incompatibility between the security work and the situation and tasks in terms of concept, system, management and technical support is becoming increasingly prominent. There is a lack of effective measures to monitor the leakage of sensitive events and control the expansion of the risk of loss of secrets. Therefore, it is an urgent task to improve the level of network security sensitive event monitoring technology and establish an effective security monitoring and management system for Internet office security management.

ITEI 2022, December 23-25, Harbin, People's Republic of China

DOI 10.4108/eai.23-12-2022.2329079

2 ANALYSIS ON RISK POINTS OF SECURITY WORK

2.1 Weak awareness of confidentiality personnel

Confidentiality awareness of secret related personnel is weak, which reflects the lack of confidentiality knowledge, the lack of understanding of the scope of documents, the lack of clear identification of secrets, and the fact that confidentiality is far away from their own behavior. Some personnel did not estimate the complexity and severity of the form of confidentiality. They thought that their work only involved work secrets, and they could not access highly sensitive documents. Therefore, they did not need to pay attention to confidentiality work. Little did they know that the leakage of state secrets and company secrets could be regarded as a loss of confidentiality event, and there was a situation of passive disclosure. The existence of the fluke mentality makes the confidentiality work more difficult. As long as we relax our vigilance for a moment, the lawbreakers will have an opportunity to take advantage of it.

2.2 Threat of online office

Network security has always been a hot topic in recent years. Various network attacks are impossible to prevent. Professional white hats cannot guarantee their own network information security, let alone the information security of ordinary people. In the process of daily online office work, the incorrect use of mobile devices, the random connection of wireless networks, and the outgoing waiting of sensitive files all pose the threat of data leakage. The loss of mobile devices is also an important threat. According to the survey results, 70 million mobile phones are lost every year, 60% of which contain sensitive information; Mobile photos expand the spread of sensitive information on the Internet; Public and private networks are mixed. Mobile terminals have both personal applications and enterprise applications. There is no obvious separation between enterprise and personal data. The support means for data tracking and review and online behavior audit are insufficient, and the risk of enterprise secret disclosure is huge.

2.3 Inadequate technical prevention means

Weak passwords still exist in office terminals. The lack of life-cycle control measures for confidential information equipment and confidential carriers leads to unclear definition of the scope of access to confidential content, which may lead to the risk of loss and disclosure. The illegal storage of sensitive files in office terminals is the source of the loss of confidentiality, and there is a lack of technical support for sensitive file monitoring.

3 RESEARCH ON SENSITIVE EVENT VISUALIZATION BASED ON SEMANTIC ANALYSIS

3.1 Technical route and system architecture

This paper studies the intelligent character recognition technology of files, pictures, etc., combining word segmentation technology, tf idf weighting technology, with the help of simHash algorithm and minHash algorithm, to effectively improve the working efficiency of

security personnel, and to turn confidential sensitive events into scenarios and processes, so as to form a visual presentation of the foreground, data collection and analysis of the background The technical route of terminal host probe monitoring.

The overall technical route adopts a three-layer structure: the terminal host probe monitoring layer, which uses technical means to scan and monitor various behaviors of illegal storage, operation and transmission of sensitive files; The background data collection and analysis layer establishes four types of data analysis scenario models to summarize, process and analyze various types of data collected by terminal monitoring; The foreground visual presentation layer presents the background analysis results to users in a centralized and real-time manner using various visual charts, and presents the data according to the user's importance level and domain, so as to achieve accurate user behavior positioning, real-time remote online automatic monitoring, large screen visual display, and comprehensively improve the awareness of confidential security sensitive events.

The overall architecture of the system mainly includes three functional modules: security inspection, real-time monitoring of sensitive information and basic information management. For administrators, it is mainly to master the practical skills of these three functional modules, while end users mainly master the common sense of self inspection and real-time monitoring information viewing of sensitive information, as shown in Figure 1.

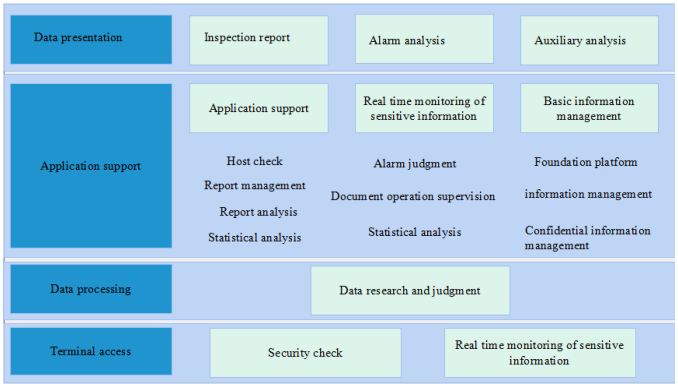

Figure 1. System architecture diagram

3.2 Research on sensing technology of sensitive information real-time monitoring

The main technical point of real-time monitoring and perception of sensitive information is to monitor and capture the behavior of user terminals when they perform various file operations, which is realized through Hook monitoring, disk underlying file scanning and other technologies.

Hook monitoring technology implementation principle: Hook technology can not only protect files and prevent files from being tampered with, but also monitor operating system processes, which can protect compliant processes and force illegal processes to close. In this paper, Hook technology is encapsulated and designed to achieve a modular Hook platform.

Implementation principle of disk bottom file scanning technology: disk bottom file scanning is a technology to identify, analyze and extract all contents of various files in the system by

analyzing file system principle, OLE document nesting mechanism and various binary formats of document files. This technology can be used to extract abnormal files such as nested hidden files, damaged files, and files with modified suffixes, so as to eliminate illegal bypass operations.

In addition, through the research on nearly 4000 kinds of nested combinations, the nested document inspection technology is implemented for the first time, which solves the problem that illegal files are nested and hidden in regular files. This technology supports multi-level nested inspection of multiple types of office documents, such as DOC, DOCX, PPT, PPTX, XLS, and XLSX. On this basis, the paper optimizes the file spot check algorithm, studies the text extraction technology, and realizes the document check of file header damage, file suffix modification, etc., making the file check complete without omission [2].

3.3 Research on intelligent decision technology based on semantic analysis

In view of the heavy workload of file determination, repeated file inspection and reporting, this paper studies multiple technologies such as OCR identification, semantic thesaurus filtering [3], fingerprint comparison and so on to achieve intelligent file determination and effectively improve the accuracy of sensitive event monitoring and determination.

- OCR identification technology. The inspection terminal realizes the inspection of picture and image files by recognizing the picture and image files and converting them to text format. First, the image is preprocessed and the imaging effect of the image is adjusted. Next, the text layout is analyzed; First, split each line to cut the characters of each line; Then, each line of text is divided into columns and finally cut into characters; Send the characters to the trained OCR recognition model for the most important character recognition, and get the final recognition results.

- Semantic thesaurus filtering. When the terminal executes the inspection task, it records the strategy of each inspection task initiated, and automatically filters the compliance words checked, such as the computer "power on password" and other phrases. Gradually accumulate, supplement and improve, build a compliance vocabulary expert knowledge base, realize automatic filtering, and reduce the system false alarm rate.

- Fingerprint comparison of similar documents. File fingerprint matching technology is to generate fingerprint feature library from sample document, and then extract fingerprint from document or content to be detected by the same method; The obtained fingerprint is matched with the fingerprint database to obtain its similarity [4]. The terminal calculates the fingerprint signature of the file through the fingerprint algorithm. After receiving the file fingerprint information, the server uses word segmentation technology, information retrieval and data mining weighting technology TF-IDF, with simHash and minHash algorithms, uses the violation files in the sensitive file library as the comparison source, finds out the specific files that meet a certain similarity threshold from the comparison source data, and realizes the auxiliary file determination function.TF-IDF weighting technology is a statistical method to evaluate the importance of a word to a file set or one of the files in a corpus. TF-IDF weighting technology is used to count the keywords in the file. The calculation method is as follows:

$$TF_n = \frac{\text{The number of times a term n appears}}{\text{The number of all terms in this class}}$$

$$IDF = \log\left(\frac{\text{The total number of documents in the vault}}{\text{The number of documents containing term n}+1}\right)$$

TFn word frequency refers to the number of occurrences of a given word in the file. The frequency of IDF reverse file refers to that the fewer documents containing the term t, the larger the IDF, which indicates that the term has good classification ability[4].

$$TF - IDF = TF * IDF$$

It can be seen from the formula that the TF-IDF weighting technique reflects that the importance of a word increases proportionally with the number of times it appears in the file, but decreases inversely with the frequency of its appearance in the corpus. In other words, if a word appears repeatedly, it will be used as a keyword regardless of its location.

- Layout file check. After analyzing, summarizing and summarizing the formats of red headed confidential documents and other relevant documents issued by government agencies, we can form unified and quantifiable indicators for such specific format documents. During the terminal inspection tool inspection, we analyze the matching degree of each indicator and then check the standard format documents, find the format documents that really meet the indicators, and display them separately on the page, Improve the attention of document operators to this type of document and assist in judgment.

- Automatic determination of the same file. When the terminal checks the file, it calculates the MD5 value of the file through the MD5 information digest algorithm, which is used as the unique identification information of the file when the report is submitted. When the communication server receives the report, it compares the MD5 value of the file with the MD5 value of the sensitive file library file, and automatically judges the current file based on the judgment result of the sensitive file library when it finds a consistency, so as to realize the automatic judgment of the file and simplify the workload of the judges.

- Application of intelligent judgment technology. According to the previous work accumulation, 1000 files were selected as experimental data. Based on the semantic thesaurus, the files are classified to ensure that the proportion of each type of file meets the requirements in Table 1 below.

TABLE 1. TEXT CHARACTERISTICS

Classification of documents required for experiment	**Cannot match keyword file**	**Matching keyword blacklist file**	**Put file library file**	**Different degrees similar to the file library file**
Proportion	5%	10%	65%	30%

First, preprocess the file, determine the keywords of the article and calculate the similarity of the file, filter out common words through calculation, and retain important words. The preprocessed file is generated into a unique MD5 file identifier to form a text information fingerprint. The TF-IDF value of words in the document is calculated [5]. With simHash and minHash algorithms, specific files that meet a certain similarity threshold are found, as shown in Figure 2.

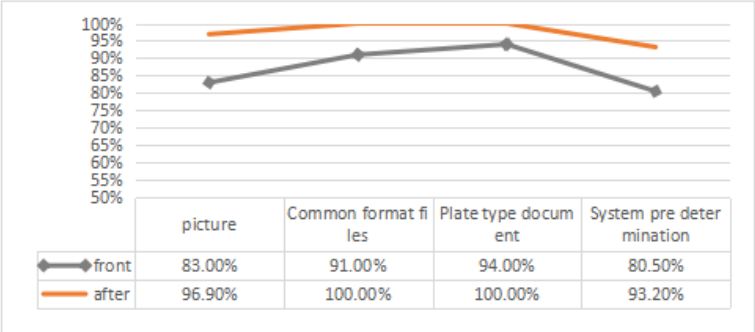

Figure 2. Recognition accuracy

The experiment shows that the accuracy rate of document determination is 93.2%. There are different forms of terminal files, and the paths of file transmission are diversified. The above intelligent technologies are comprehensively used to realize the comprehensive monitoring of static files and dynamic files. At this stage, it still requires manual secondary judgment to ensure the accuracy of the judgment results, but it has achieved the goal of greatly improving the level of network security sensitive event monitoring technology and saving manpower.

3.4 Visual monitoring platform for sensitive events

Based on the above technical research, a set of network security sensitive event monitoring platform is designed and implemented, which enables online real-time scanning of office terminals to monitor various behaviors of illegal storage, operation and transmission of sensitive files. In the aspect of display module design, the user's importance level, domain and time sharing are used for display, multiple data collection and analysis scenario models are established, various data collected by terminal monitoring are summarized and processed, and the analysis results are displayed in a diversified manner through rich components.

In terms of the visual presentation of sensitive events, due to the large amount of data monitoring the operation behavior of terminals throughout the network, and considering the response speed and display effect of the comprehensive display platform, the mainstream frameworks Angular, WebGL three-dimensional technology, grid technology, etc., which are more advanced in the front end, are adopted to ensure the perfect presentation of visual effects.

4 CONCLUSIONS

"There is no national security without network security". Network informatization is a double-edged sword. While promoting rapid economic development, it also brings more risks and challenges to security. Based on the requirements of enterprise level security work, this paper researches and implements the semantic analysis technology, and applies it to the security sensitive event monitoring business, so as to achieve intelligent perception and visual display of sensitive events, realize the scientific operation and management of enterprise level security work, and effectively improve the comprehensive defense capability of enterprise security. Next, we will continue to improve the semantic analysis knowledge base, improve the

accuracy of machine learning algorithms, expand the visual presentation of sensitive events, enrich monitoring scenarios, and gradually promote the risk perception monitoring of sensitive events on mobile terminals to achieve comprehensive control of sensitive information and strictly prevent the occurrence of leakage events.

REFERENCES

[1] Wang Ping, Zhou Wei, Jia Dan, Li Jie & Lu Zhiqiang. (2022). Deepen the research and application of security management system to realize the transformation and upgrade of security work. Space Industry Management (09), 95-101.

[2] Duan Lijuan, Zhang Xiqun, Ma Longlong & Wu Jian.(2017).Text extraction method for historical Tibetan document images based on block projections. Optoelectronics Letters (06),457-461.

[3] Xue Yi, Li Zheng Han, Wang Bin, Liu Yunpeng, Sun Dong, Wu Fei fan. & Wang Sheng. (2022). Text semantic analysis system based on convolutional neural network. Information recording material (04), 112-114. Doi: 10.16009/J. CNKI. CN13-1295/T.Q. 2022.04.017.

[4] Liu Wenlong. (2015). Research on key techniques of digital fingerprinting (master's thesis, Beijing University of Posts and Telecommunications). Https://kns.cnki.net/kcms/detail/detail.aspx?dbname=cmfd201502&filename=1015583643.nh

[5] song Lili. (2012) . Research on fingerprint identification technology and its application in file encryption system (master thesis, Hebei University of Science and technology). https://kns.cnki.net/kcms/detail/detail.aspx?dbname=cmfd201302&filename=1012337575.nh