Research on Leakage Prevention Technology of Sensitive Data based on Artificial Intelligence

Donglan Liu[1*], Xin Liu[1], Lei Ma[1], Yingxian Chang[2], Rui Wang[1], Hao Zhang[1], Hao Yu[1], Wenting Wang[1]
[1]*State Grid Shandong Electric Power Research Institute, Jinan 250003, Shandong Province, PR China*
[2]*State Grid Shandong Electric Power Company, Jinan 250000, Shandong, PR China*
*E-mail: liudonglan2006@126.com

Abstract— **In recent years, network security incidents happen frequently, and each security incident is inseparable from the threat of data leakage. In order to deal with the sensitive data leakage under the new network threat, this paper proposes the sensitive data leakage prevention method based on artificial intelligence technology by studying the current data leakage prevention technology. By prefabricating a simple security policy, the detection result of the data and the data are submitted to the security manager. According to the judgment of the security inspector, the self-learning security policy is carried out with artificial intelligence algorithm to promote the update of the security policy. Data assets leakage protection system is designed to show full view data assets directory, verify the risk analysis technology in data assets for centralized cataloging and grade of risk distribution data visualization display, based on the classification properties of the data security label of the multi-function scenarios such as the technology application effect, effectively reduce the risk of data leakage.**

Keywords- artificial intelligence; sensitive data; data leakage prevention technology; data asset leak prevention system

I. INTRODUCTION

With the rapid development of information technology, the idea that data is the core asset of government and enterprises has been widely accepted. As a result, how to protect data assets has become the top priority of data security. The electric power industry is the basic industry of the national economy and one of the extremely important infrastructures for the development of the national economy and people's life [1-2]. In recent years, under the background of smart grid and global energy Internet, the grid information technology keeps developing, and more and more data of various smart grid service systems are stored centrally [3-4]. Business system data is an important asset of the organization. It is not only valuable to the business of smart power grid distribution, transmission, transformation, distribution and electricity consumption, but also of great significance to the decision-making and strategic planning of power grid companies. At the same time, China's information system security vulnerabilities continue to emerge, the network security situation is very serious, information security events emerge in an endless stream, the power industry business system is also facing severe security threats [5].

In the process of power grid construction and operation, a large amount of data information is generated. Once these data information is leaked, the security of power users' personal information will be disclosed, causing huge economic losses to power users and power companies. Therefore, the application of data leakage prevention technology in the operation process of power grid informatization construction can effectively prevent the leakage of data information and ensure the safety of power grid operation data information.

At present, domestic and foreign network security protection and data leakage prevention technology has certain development. Zhao yong et al. [6] proposed a defense model of information leakage in enterprise Intranet security, which is based on password isolation and uses access control and password technology to build a virtual secret-related network in the enterprise Intranet to prevent the leakage of sensitive data. At the same time, the research and construction of data leakage prevention and information security protection system are carried out by various industries in combination with the needs of leakage prevention of sensitive data of enterprises [7-14].

In order to deal with the sensitive data leakage under the new network threat, a sensitive data leakage prevention method based on artificial intelligence technology is proposed, and a data asset leakage prevention system is designed to cope with the complex and changeable network threat situation.

II. RESEARCH ON DATA LEAKAGE PREVENTION TECHNOLOGY

A. Deep Content Recognition Technology

Deep content recognition technology is mainly based on keyword, regular expression, structured data fingerprint, document fingerprint, machine learning and other technologies.

1) Document fingerprint detection

Document fingerprint detection technology is the generation of unstructured data fingerprint to form sensitive data fingerprint characteristic library. That is, after the fingerprint is generated for the protected document content, it is matched with the fingerprint of the outbound document block. If the fingerprint coincidence degree exceeds a certain threshold, the contents of the outbound sensitive document can be judged.

Document fingerprinting technology is used to protect unstructured data in the form of Microsoft Word and

Foundation item: Scientific Research Program of State Grid Shandong Electric Power Company, Project Name: Research on Key Technologies of Network Security Protection in Smart Grid - Topic 3: Research on Key Technologies of Data Security Analysis and Privacy Protection for Big Data, ERP Number: 520626200013, Contract No. 2020A-015.

978-1-7281-6313-0/20/$31.00©2020 IEEE

PowerPoint files, PDF documents, source files, CAD/CAM images, and more. Document fingerprints are created to detect excerpts from source documents, drafts, or different versions of protected documents, as well as exact matches with binary content. To enforce unstructured document matching, you first provide a set of documents that contain the specific content that the unit wants to protect. These documents are then fingerprinted using the administrative console to form a fingerprint library, and unstructured data detection rules are configured to detect protected documents.

2) Structured data fingerprint

A structured fingerprint can protect customer and employee data, as well as other structured data typically stored in a database. For example, a customer might write a policy about using structured data fingerprinting to look for any three of these simultaneous occurrences of name, id number, bank account number, or phone number in a message and map them to a record in the customer database.

The structured fingerprint is usually used to construct the entity of the structured data, which can be used to protect the structured data. Usually, an independent fingerprint is constructed for each cell, and a multi-dimensional multi-level tree structure (B+ tree or R tree) is used to complete the index construction of data fingerprint. The index structure can speed up the retrieval decision of data.

3) Machine learning feature extraction and detection

Support Vector Machines (SVM) is based on the VC dimension theory of statistical learning theory and the principle of structural risk minimization. The information provided by limited samples is used to seek the best compromise between the complexity of the model and the learning ability, so as to obtain the best generalization ability. The basic idea of machine learning is to map training data nonlinearly into a higher-dimensional feature space (Hilbert space), in which a hyperplane is found to maximize the isolation edge between positive and negative examples. It shows many remarkable properties in machine learning problems such as small sample, nonlinearity and high dimensional data, and is widely used in pattern recognition, data mining and other fields.

Machine learning usually uses supervised learning algorithm to make judgment on the category (important/unimportant) of data. Algorithms commonly used include KNN, SVM, Boosting, etc. By learning the provided sample document data, extract the key features of the document data, build the minimum loss function, obtain the best classification hyperplane function, through which to complete the determination of new data.

B. Defects of Existing Data Leakage Prevention Technologies

Based on personal privacy risks and characteristics of different business scenarios, anti-leakage methods of various types of private data are determined. Existing solutions to prevent leakage of sensitive data usually use predefined security policies, such as setting key words of important data, regular expressions of sensitive information, even data fingerprints, machine learning methods based on classification, etc. These rules need to be preprogrammed. The main drawbacks of prior art are as follows.

Firstly, security policies need to be prefabricated. In most cases, enterprises cannot clearly know all the important and sensitive data, so most of the prefabricated policies are incomplete.

Secondly, the security policy is not updated in a timely manner. Even if sufficient prefabricated security policy is adopted, with the business development of the enterprise, new data will be generated every day, and the security policy cannot completely cover the newly generated data. And over time, some old data is no longer important or sensitive, and no need to be protected.

Thirdly, security policy needs to be constructed by security personnel, and only when security personnel fully understand the distribution of data can a complete security policy be constructed. Some data will not be used, while some data will be used frequently. Security personnel cannot know which data should be protected. Security personnel have a lot of work to do.

Fourthly, the prefabricated security policy will misjudge the data (judge the non-important data as important data) in the actual application process, which increases the difficulty of data leakage prevention. If the security policy is adjusted, it can lead to missed decisions (important data is judged as non-important data).

III. SENSITIVE DATA LEAKAGE PREVENTION TECHNOLOGY BASED ON ARTIFICIAL INTELLIGENCE TECHNOLOGY

A. Principle of Sensitive Data Leakage Prevention Technology Based on Artificial Intelligence

The sensitive data leakage prevention technology based on artificial intelligence technology prefabricates a simple security policy and presents the detection results of the data together with the data to the security manager. According to the actual business situation and the scene where the data appears, the security manager gives the judgment of whether the test result is correct or not. If the test result is correct, a positive score is given. If the test result is wrong, a negative score is given. According to the judgment of security detection personnel, artificial intelligence algorithm is used for self-learning, and the learning result is set as security policy, which is effective for subsequent detection.

By building a measurement model for the importance (also known as sensitivity) of the document. Promote security policy updates by establishing a uniform measure of document importance, measuring the importance of documents, presenting data clearly to managers.

The realization process of sensitive data leakage prevention based on AI technology is as follows.

Algorithm 1: Sensitive data leakage prevention technology based on AI technology

1. Enable prefabricated security policies.

2. Initialize the detection model. Detection models are not limited to neural networks, but also include intelligent algorithms such as machine learning.
3. The prefabricated security policy is used to detect the data and obtain the detection results.
4. Report the test results and data, output the feedback score results.
5. Receive scores of test results and data.
6. According to the detection results, data and scores, the detection model adopts batch gradient descent algorithm to obtain the optimal solution of the detection model, and then generates a new security detection strategy.
7. The new security detection strategy is applied to the data leakage prevention system to detect the subsequent data.
8. Repeat steps 4 through 9, and the security policy can be continually updated.

Algorithm 2: Document comparison algorithm

1. Content extraction is carried out for documents in the network data leak prevention system.
2. Perform lexical analysis and syntactic analysis on document contents, such as using HMM for lexical analysis, using dependency tree for syntax analysis, and eliminating stop words, such as "of".
3. The maximum likelihood estimation algorithm is used to calculate the probability of the word.
4. Information entropy algorithm is used to obtain entropy value for each document.
5. Normalize the entropy.
6. The normalized entropy value is taken as the important degree value of the document.

Generating a security policy based on the importance of the document, such as building a fingerprint on sensitive data for security policy, or extracting important keywords from sensitive data for security policy.

B. Data Asset Leakage Prevention System Design

Data asset leakage prevention system is designed to prevent and detect data asset leakage of diversity, heterogeneity and application complexity in big data environment. The existing data storage and management methods are single, so it is difficult to prevent data leakage effectively. Data leakage prevention system and its security detection tools are classified according to the data classification and grade index, data fingerprint extraction and other characteristics. According to the company's data leakage protection strategy, the enterprise unified construction of the Internet export data content audit. Through the company's important data or information assets in violation of the security policy in the form of the behavior of the outflow of the company security detection, so as to effectively reduce the leakage of the company's important data assets.

The data asset leak prevention system designed in this paper can display the data asset catalog from the whole perspective. The system can verify the application effect of risk analysis technology in multi-functional scenarios such as centralized cataloging of data assets, visual display of data risk distribution, and data security label technology based on classification and grade attributes.

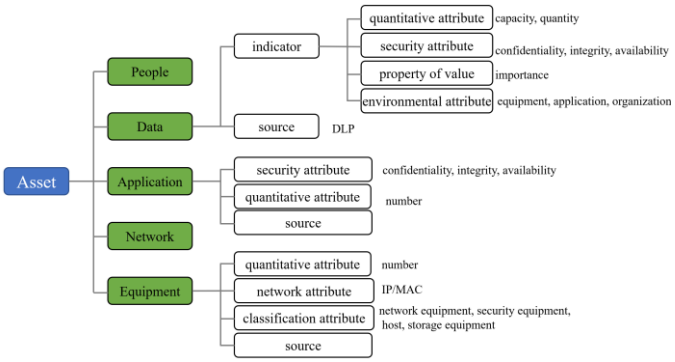

Figure 1. " Key elements of risk assessment for data assets in the prototype system.

As shown in figure 1, the main capability of risk assessment of data assets will be realized in the prototype system, and necessary data and information will be collected from people, data, applications, network and equipment, etc. Finally, comprehensive analysis is carried out. The specific element information collected is shown in figure 2 and figure 3.

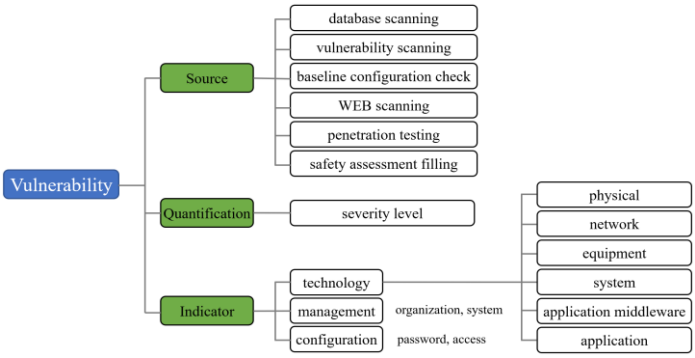

Figure 2. " Key elements of vulnerability assessment of data assets in prototype systems.

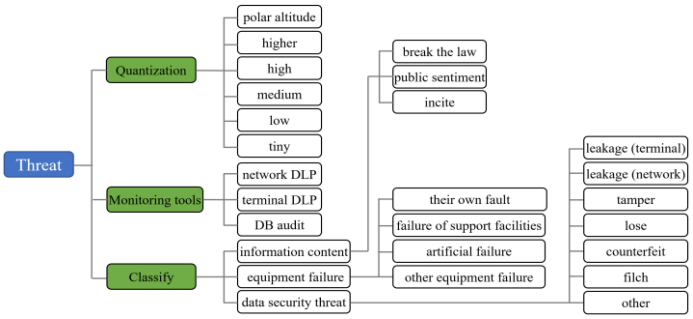

Figure 3. " Key elements of the prototype system's data threat assessment.

The prototype system will implement the following capabilities. The data asset view is shown in figure 4. The data asset risk view is shown in figure 5.

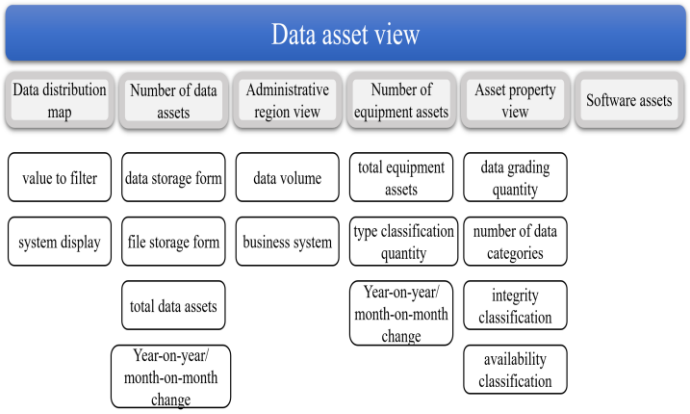

Figure 4. " Prototype system data asset view capability content.

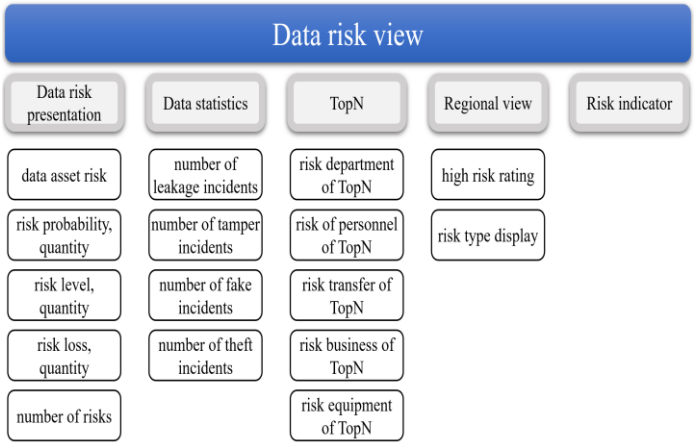

Figure 5. " Prototype system data asset risk view capability content.

Through the design of the data asset leak prevention system to show the full perspective of the data asset directory. The system will realize the security metadata management, the data assets centralized catalog, the data distribution view display, the data classification and grade attribute and the security label setting ability. And the realization of data label based control technology, automatically identify the sensitivity of data. In practice, label control technology can be combined with other data security technology. For example, based on security metadata, different levels of dynamic and static desensitization methods are adopted for data of different types, formats, classifications and security levels.

IV." CONCLUSION

In this paper, artificial intelligence technology innovation is applied to the field of sensitive data leakage prevention. Through the use of deep content recognition and machine learning and other technologies to achieve the user business system core and sensitive data collection mining and accurate identification. And the realization of content features real-time analysis and data visualization. The accurate data security model is formed by generating protection strategies from the sensitive information samples. When the occurrence of sensitive information violation outward behavior, the system can record, alarm and block the violation of sensitive information. This can maximize the security of sensitive information storage and use and reduce the risk of data leakage.

ACKNOWLEDGMENT

This research was financially supported by the State Grid Shandong Electric Power Company (No.2020A-015). Project Name: Research on Key Technologies of Network Security Protection in Smart Grid - Topic 3: Research on Key Technologies of Data Security Analysis and Privacy Protection for Big Data, ERP Number: 520626200013. Meanwhile, the authors wish to acknowledge the anonymous reviewers and all co-workers for their helpful comments.

REFERENCES

[1]" WANG Dong, CHEN Chuanpeng, YAN Jia, et al. Pondering a new-generation security architecture model for power information network[J]. Automation of Electric Power Systems, 2016, 40(2): 6-11.

[2]" YU Yong, LIN Weimin. Design of the electric power system's security monitoring system based on classified protection[J]. Computer Science, 2012, 39(11A): 440-442.

[3]" DENG Song, LIN Weimin, ZHANG Tao, et al. Applied planning of data leak prevention in power grid information construction[J]. Electric Power Information Technology, 2013, 11(1): 1-4.

[4]" JIANG Chengzhi, YU Yong, LIN Weimin. Research on electric information network security situation awareness model based on intelligent agent[J]. Computer Science, 2012, 39(12): 98-101.

[5]" CHEN Hong. Research on the Grid Sensitive Data Comprehensive Defense System Based on Dual Defense of Technology and Management[J]. Electric Power Information and Communication Technology, 2016,14(9):42-45.

[6]" ZHAO Yong, LIU Jiqiang, HAN Zhen. The application of information leakage defendable model in enterprise intranet security[J]. Journal of Computer Research and Development, 2007, 44(5): 761-767.

[7]" CAI Xi-ping, LUAN Yuhang, TANG Shen. Addressing internal network threats and corresponding counter measures for enterprise network[J]. Network Security Technology & Application, 2007(3): 65-67.

[8]" CHANG Runmei, MENG Liqing. Study on sensitive data vault management of telecom industry[J]. Information Security and Communications Privacy, 2013(8): 82-84.

[9]" WU Zejun. Protection of enterprise data security with data leak prevention [J]. Computer Security, 2010(1): 81, 82-85.

[10]" LI Weiwei, ZHANG Tao, LIN Weimin, et al. Research and implementation of sensitive data identification method based on text content[J]. Computer Engineering and Design, 2013, 34(4):1202-1206.

[11]" TIAN Jing. Leakage Protection Technology Research of Public Security Sensitive Data under Cloud Computing Environment[J]. Journal of LiaoNing police College, 2016, 99 (5): 62-66.

[12]" WANG Wei, TIAN Bing, LIU Ying, et al. Hierarchical Scheduling Algorithm of Large Data for Distributed Systems[J], Shandong Electric Power, 2017,44(6):45-48.

[13]" LIU Donglan, SHI Fangfang, LIU Xin, et al. Research on Protection for the Database Security Based on theCloud in Big Data Environment[J]. Shandong Electric Power, 2017,44(6):41-44,48.

[14]" JIA Yujian, SUN Shumin, MEN Yu, et al. A Method for Distribution Network Fault Monitoring Based on Big Data Analysis[J].Shandong Electric Power, 2017,44(10):1-5.