Hindawi
Journal of Sensors
Volume 2020, Article ID 7358692, 9 pages
https://doi.org/10.1155/2020/7358692

Research Article

Construction of a Security Vulnerability Identification System Based on Machine Learning

Kebin Shi,[1] Yonghui Dai ,[2] and Jing Xu[3]

[1]*Advisory Department, Shanghai Information Investment Consulting Co., Ltd., Shanghai 200081, China*
[2]*Management School, Shanghai University of International Business and Economics, Shanghai 201620, China*
[3]*School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China*

Correspondence should be addressed to Yonghui Dai; dyh822@163.com

Received 19 February 2020; Revised 8 July 2020; Accepted 20 July 2020; Published 6 August 2020

Academic Editor: Fei Yu

In recent years, the frequent outbreak of information security incidents caused by information security vulnerabilities has brought huge losses to countries and enterprises. Therefore, the research related to information security vulnerability has attracted many scholars, especially the research on the identification of information security vulnerabilities. Although some organizations have established information description databases for information security vulnerabilities, the differences in their descriptions and understandings of vulnerabilities have increased the difficulty of information security precautions. This paper studies the construction of a security vulnerability identification system, summarizes the system requirements, and establishes a vulnerability text classifier based on machine learning. It introduces the word segmentation, feature extraction, classification, and verification processing of vulnerability description text. The contribution of this paper is mainly in two aspects: One is to standardize the unified description of vulnerability information, which lays a solid foundation for vulnerability analysis. The other is to explore the research methods of a vulnerability identification system for information security and establish a vulnerability text classifier based on machine learning, which can provide reference for the research of similar systems in the future.

1. Introduction

With the rapid development of the Internet, various applications and information systems based on the Internet bring convenience and efficiency to individuals and enterprises. At the same time, it also brings a lot of information security problems [1]. According to a report released by the National Computer Virus Emergency Response Center of China, a total of 7,478,639 viruses were found in December 2018, and the main transmission routes of the virus were a phishing website, Trojan virus, and information security vulnerabilities [2]. In essence, information security refers to the defects of software. Once these defects are found and used by attackers, it is very easy to cause information theft and leakage and system damage, which often leads to huge losses [3]. For example, spectre and meltdown hacks attack behavior caused by CPU chip vulnerability [4], which involves many smartphones, personal computers, and servers [5]. As an important guarantee for the stable operation of the system, information security covers a wide range of security protocols, such as SSH, SSL, and set [6], and network security authentication mechanisms such as digital authentication, digital signature, digital time stamp, and computer security operating system [7]. Because the software itself inevitably has defects, it causes security vulnerabilities. How to detect and prevent security vulnerabilities has always been one of the research hotspots in the field of information security.

In recent years, countries all over the world pay more and more attention to security vulnerabilities. A large number of security vulnerabilities have been found, which makes vulnerability management face many problems to be solved, such as the judgment and processing of vulnerability redundant data. The direct manifestation is that although there are many vulnerabilities, these vulnerabilities have the same characteristics and should be classified as similar vulnerabilities. However, the lack of a unified vulnerability identification rule makes the above vulnerabilities exist in the vulnerability database and causes vulnerability redundancy. In addition, the

correlation analysis of vulnerability, description, and identification of vulnerabilities are important contents in vulnerability management, which is very important for the construction of information security.

The remaining parts of this article are organized as follows: In Section 2, literature review is introduced. In Section 3, the vulnerability classification and machine learning method are shown. In Section 4, the system framework and design are introduced. In Section 5, the key technologies of TextRank keyword extraction and attribute extraction based on frequency are shown. Section 6 is the conclusion and discussion of this article.

2. Literature Review

2.1. Vulnerability Library and Detection Technology. At present, some security service organizations around the world have established their own vulnerability libraries. Common network vulnerability libraries mainly include CVE (Common Vulnerabilities and Exposures), NVD (US National Vulnerability Database), and CNVD (China National Vulnerability Database) [8]. Among them, each CVE vulnerability has a unique name corresponding to the CVE dictionary, which helps users distinguish vulnerabilities from vulnerability databases and detection tools [9]. If the vulnerability in the information security monitoring report belongs to a vulnerability in the CVE database table, then the corresponding patch solution can be obtained through the vulnerability name to solve the information security problem in time. For example, a CVE vulnerability number is CVE-2008-1046. NVD refers to the United States National Vulnerability Database [10]; its description of vulnerabilities includes 15 attributes such as vulnerability number, release date, vulnerability description, hazard type, attack path, and vulnerability type. It is widely used in global information security vulnerability services. CNVD is China's national information security vulnerability sharing platform. As China's official vulnerability release and security early warning platform, it plays an important role in the basic service of China's information security. The number of vulnerabilities released by it is CNVD-2019-0282. In addition, some security protection companies in China have also invested in the construction of vulnerability information resource libraries, such as the sky mirror vulnerability information resource library, which is a resource library for the announcement and protection of computer security vulnerabilities established by the company. It collected a variety of vulnerabilities and protection tools to provide users with vulnerability detection, patching, and attack verification, which improved the level of forensics and verification [11].

From the perspective of time, information security vulnerabilities often show the characteristics of the time life cycle. It will go through the process of creating and dying out of vulnerabilities. In the above process, the vulnerability presents phase characteristics. For this reason, some scholars divide the characteristics of information security vulnerabilities according to time, such as 0-day vulnerability and 1-day vulnerability [12]. In the research of vulnerability technology, more representative technologies mainly include APT detection technology, big data-based network vulnerability scanning technology, clustered vulnerability analysis technology, and intelligent vulnerability mining technology. At present, network security vulnerability detection systems include both paid commercial systems and open-source free leak scan systems, such as 360 security guards in China, Norton, Avast, and IBM Rational AppScan. APT detection and defense are an important content of information security. Chinese scholars have conducted research on user behavior and network traffic and applied social engineering to propose a baseline-based APT detection, which has traced and confirmed the APT attack [13]. Some scholars have proposed a method for predicting intrusion detection events based on APT and the functional configuration of the method, and they implemented a prediction model based on intrusion detection events through testing at the stages of learning, prediction, and evaluation [14]. With the gradual increase in software types and the development of information technology, it will become a trend to conduct large-scale vulnerability detection based on big data, artificial intelligence, and machine learning technologies in the future.

2.2. Review of Machine Learning in Network Security. In recent years, with the development of big data and artificial intelligence technology, recognition based on machine learning has been widely used in data analysis. In essence, machine learning is to simulate human learning behavior through computers. Experience is used as input, through continuous learning and iteration to train and build learning behavior models, so as to meet human-like standards for identification and prediction [15]. The common algorithms of machine learning include a regression algorithm, association rule, support vector machine, clustering algorithm, decision tree algorithm, artificial neural network, and deep learning. Some scholars have analyzed and compared the application technologies of machine learning in software defect finding, malicious code detection, and intrusion detection, including linear discriminant analysis, decision tree analysis, multiple linear regression, rough set, support vector machine, and artificial neural network [16].

The application of a machine learning algorithm in network security includes security intrusion detection, spam detection, and domain name detection [17]. For example, through the use of a random forest algorithm and SVM support vector machine algorithm, the Chinese scholars take the KDD Cup 99 data set as the sample for intrusion detection simulation analysis and get the data of the above algorithm on the false alarm rate, training time, model memory occupation, and unknown attack detection ability, as well as the advantages and disadvantages of each algorithm [18]. Some scholars use Weka and RapidMiner to evaluate the performance of a machine learning algorithm for spam detection on Twitter [19]. Their research results provide a reference for antispam. For better monitoring, some scholars put forward the MLH-IDS machine learning framework, which consists of three layers: supervised learning layer, unsupervised learning layer, and outlier detection layer. The advantages of different machine learning methods are comprehensively described, so that the framework can show more flexibility

and good performance [20]. In addition, some scholars layered the data, combined the SVM support vector machine and the ELM algorithm, and built a multilayer hybrid intrusion detection model. The model was tested on the KDCUP99 data set and achieved an accuracy of 95.75% [21].

3. Theory and Technology

3.1. Vulnerability Classification. Vulnerability classification refers to the classification of vulnerabilities. From the perspective of mathematical thinking, the vulnerability classification process is a mapping process. It classifies vulnerabilities that need to be classified into existing vulnerability categories according to a certain mapping relationship. The vulnerability category refers to the type of vulnerability, which is divided into categories based on attributes such as the cause of the vulnerability, the scope of action, the technology used, and the location characteristics.

There are many forms of information security vulnerability and its deformation. Therefore, many countries have established a special information security vulnerability database. As an important force of information security maintenance in the century, China has established the China National Vulnerability Database of Information Security (CNNVD), which comprehensively describes the classification of vulnerabilities. It mainly includes general vulnerability, event vulnerability, and public vulnerability. Specifically, it divides vulnerabilities into 26 categories, including configuration errors, input verification, code problems, and SQL injection. A sample of vulnerability types is shown in Figure 1 [22].

It can be seen from Figure 1 that the first level of vulnerability is divided into three major categories, namely, configuration errors, code problems, and insufficient information. Among them, the code problem category can be divided into sublevel categories. In the above vulnerabilities, configuration error vulnerability refers to the vulnerability in the process of software configuration, which is caused by unreasonable configuration in the use of the software.

3.2. Machine Learning. Machine learning, literally, means to provide some data to personal computers, servers, and other machines and let them learn and find out the logic of data through mathematical modeling and self-iterative method, and then, it can automatically complete prediction, classification, and recognition once it faced similar data. At present, machine learning has been widely used in pattern recognition, visual visualization, and network intrusion detection and other fields [23]. From the perspective of the application of machine learning, it can be divided into five categories: supervised learning, semisupervised learning, unsupervised learning, transfer learning, and reinforcement learning [24]. The complete process of machine learning consists of business understanding, data collection, data preprocessing, data modeling, and model evaluation [25]. Among them, business understanding refers to understanding the needs and background knowledge of the task before performing the task of machine learning. The data collection mainly includes the collection and storage of the original data, which is the premise of follow-up work and provides the basis for future work. Data preprocessing is to clean and transform the original data, which is a very important process in machine learning. In this process, effective information needs to be extracted as much as possible to prepare for subsequent modeling. Data modeling refers to the establishment of a data model by machine learning methods such as supervised learning, unsupervised learning, and reinforcement learning. The classic supervised learning algorithms include artificial neural networks, naive Bayes, and decision tree analysis, and the classical unsupervised learning algorithms include clustering and dimension reduction. Model evaluation refers to the use of some relevant methods and indicators to evaluate the advantages and disadvantages of the model obtained by a machine learning algorithm, and its common evaluation indexes include precision, recall, F-value, and accuracy.

SVM (support vector machine) is a typical algorithm in machine learning. Its core idea is to find the most suitable separation hypersurface in the sample space, which can distinguish the samples significantly. Common forms of SVM include linear separable, linear support, and nonlinear support vector machines. Among them, the linear regression of SVM is expressed as follows.

Set the sample set as $(y_1, x_1), \cdots, (y_l, x_l), x \in R^n, y \in R$, and use a linear equation to represent the regression function.

$$f(x) = \omega^T \varphi(x) + b. \tag{1}$$

The essence of formula (1) can be regarded as a constrained optimization problem, and its expression is as follows.

$$\varphi(\omega, \xi, b) = \frac{1}{2}|\omega|^2 + C\left(\sum_{i=1}^{l} \xi_i + \sum_{i=1}^{l} \xi_i^*\right). \tag{2}$$

In formula (2), C refers to the penalty factor and ξ and ξ^* represent the upper and lower limits of the relaxation variable, respectively. Formula (2) is solved by the Lagrangian constraint equation, which is shown as follows.

$$\bar{\alpha}, \bar{\alpha}^* = \arg\min \left\{ \begin{array}{l} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right)\left(\varphi(x_i)\varphi\left(x_j\right)\right) \\ \quad -\sum_{i}^{l}(\alpha_i - \alpha_i^*)y_i + \sum_{i}^{l}(\alpha_i + \alpha_i^*)\varepsilon \end{array} \right\}, \tag{3}$$

In formula (3), $\varphi(x)$ is a kernel function. If $\varphi(x_i)\varphi(x_j) = x_i x_j$, then it represents a linear support vector machine; otherwise, it is a nonlinear support vector machine. The solution expressions of the sum of the coefficients to be

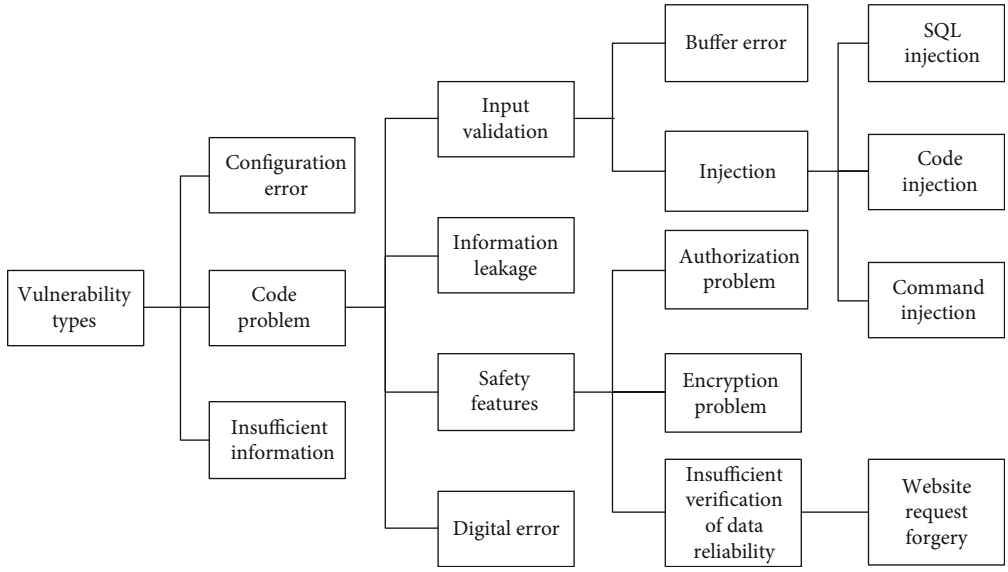

Figure 1: Sample of vulnerability types.

determined, the regression coefficients, and the constant terms are as follows.

$$\begin{aligned} \bar{\omega} &= \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) x_i, \\ \bar{b} &= -\frac{1}{2} \bar{\omega} [x_r + x_s]. \end{aligned} \tag{4}$$

In a kernel function, the Gaussian kernel is widely used. It is also called radial basis kernel (RBF kernel), and its expression is as follows:

$$\varphi(x_i)\varphi(x_j) = \exp\left(-\frac{|x_i - x_j|^2}{2r^2}\right). \tag{5}$$

In formula (5), r refers to the variance of the Gaussian function in the Gaussian kernel function.

4. System Design

4.1. System Design Objectives. A security vulnerability identification system is an important part of information security construction. The design of the system based on machine learning is of great significance for timely and accurate determination of a vulnerability category and hazard level and provision of a complete data source for researchers. On the basis of full investigation, the principles of system objective design are as follows.

(1) The system needs to follow the principles of flexibility and scalability. It should have good scalability and meet the needs of future upgrades

(2) The system needs to follow the principle of adaptability, and it should have environmental adaptability and consider the needs of users in order to adapt to various user operations

(3) The system should have the function of authority grading. In the system, the access control rights of users with different permission levels are different, and it is necessary to strictly control the user's rights of adding, deleting, modifying, and searching

(4) The system should follow the stability principle and consider the safety and stability of operation. It should have corresponding measures in antivirus attack, data backup, and data recovery

(5) The system design should follow the principle of modularization and divide the function of the system reasonably. For example, it has the data preprocessing module, which is responsible for cleaning the dirty data and processing the distorted data or integrated data. It has the function of vulnerability rule management, which is responsible for the addition, deletion, modification, and query of rules. It has the function of vulnerability data identification and classification

4.2. System Requirement Analysis. The purpose of the security vulnerability identification system is to integrate and uniformly manage the data and information of various vulnerability databases. It can collect, summarize, and display vulnerability data. After investigation and analysis of the current situation, the requirements of the security vulnerability identification system are summarized as follows.

4.2.1. Requirements for Establishing a Unified Vulnerability Database Description. A vulnerability knowledge database is an important accumulation of current vulnerability knowledge in the field of network security. For different vulnerability databases, there may be the same vulnerability information,

but the rules of the vulnerability number are all customized, which increases the difficulty of vulnerability knowledge management. There are many security vulnerability databases domestically and internationally such as vulnerability databases of CVE, CNNVD, and CNVD. Different security vulnerability databases have different naming and numbering rules. For example, the typical numbering method of CVE vulnerability databases is "cve-2014-4664," which is a vulnerability number. Therefore, the current main numbering rule is vulnerability database name plus discovery year plus vulnerability sequence number. Another example is a vulnerability identified as "cnvd-2017-17486" in the CNVD vulnerability database. This vulnerability belongs to buffer overflow vulnerability, which occurs in basic applications such as a database and causes database service interruption. The scope and harm of this vulnerability will be relatively wide. However, the vulnerability "cnvd-2017-17486" was found by a company in China. There is no corresponding vulnerability number in CVE. In order to better conduct vulnerability analysis, it is necessary to form a relatively unified vulnerability database.

4.2.2. Requirements for Data Quality. Because data analysis is based on a single data set, when there is data from multiple data sets, a unified preprocessing of the multiple data is required to improve the quality of the data so that the subsequent analysis can be better performed [28]. In order to improve the data quality, this paper establishes the identification management and rule-making. After the original data is obtained, the original data needs to be effectively sorted out according to the identification before it can be used for subsequent analysis. Usually, the first step is to standardize the data. Since the data processing is based on the identification data, the efficiency of data analysis and data tracing can be greatly improved.

At the same time, in order to improve the comprehensive utilization efficiency of all kinds of data, it is necessary to collect, process, and integrate the data of different data providers. In this paper, the open interface is used to manage the vulnerability rules, and the collected data is sorted into preprocessing data according to the specified format. The system takes identification rules as an important process of basic data processing and provides services for rule management, data analysis, and presentation of vulnerabilities, so as to realize centralized control and analysis of various data, especially to control data duplication and distortion and improve data quality.

4.3. System Framework. According to the requirements of the system, the system is divided into four layers: business presentation layer, application system layer, application support layer, and data resource layer. The system framework is shown in Figure 2.

The business presentation layer displays the business functions of each application management module according to the position and authority of the login personnel, such as the authority of adding, deleting, modifying, and querying the rule management function module.

The application system layer mainly relies on the various services provided by the application support layer and provides users with application software modules of the vulnerability security system according to the actual needs, including rule management, task management, data processing, result query, and data display functions.

The application support layer completes the interface services and management services related to system functions, including unified data interface, message middleware, distributed storage management, and security log audit, as well as various basic information online query and comprehensive query services and comprehensive analysis services.

The data resource layer encapsulates the data-related content, including text file, Excel file, XML file, and JSON file.

5. Key Technologies

5.1. TextRank Keyword Extraction. In the research of text classification, the support vector machine algorithm has good generalization ability, and it has a significant effect on small sample nonlinear classification. Therefore, this paper uses a support vector machine to achieve text classification. The process of text classification is complicated, mainly including text preprocessing, feature selection, classifier selection, and performance evaluation, which is shown in Figure 3.

It can be seen from Figure 3 that the process of classifying vulnerability description is as follows: firstly, text preprocessing is carried out for training data, including text segmentation, and a text model is used for characterization after removing stop words with little classification significance such as punctuation marks and characteristic characters; then, feature selection is carried out for the text model and features matching weight are selected, and then, a classification model is constructed and use the test data set to evaluate the performance of the classification model. If it meets the requirements, the classifier is selected for text classification. For example, the following vulnerability description information is classified and implemented as shown in Figure 4.

After the segmentation, the interjections, auxiliary verbs, and conjunctions are removed, and then, the characteristic words such as "SQL, database, deception, server, and malice" are selected for model training and output. In SVM classification, the choice of penalty parameter C and kernel parameter g is closely related to the performance of the SVM classifier. They control the empirical risk and VC confidence, respectively. In this study, the penalty parameter $C = 80.1532$ and kernel parameter $g = 0.23$ are obtained by continuous optimization of the kernel function.

TextRank's idea of keyword extraction is based on PageRank's idea. PageRank, as its name implies, ranks the importance of web pages. Its core idea has two main points. One is that if a web page has a large number of links with other web pages, it means that the importance of this web page is relatively high, and its PageRank value is relatively large; the other is that if a web page with a large PageRank value is compared with another page that has links, the PageRank value of the page connected with this large PageRank value will be increased accordingly. The advantage of PageRank is that the PageRank value of all its pages can be calculated statistically offline, but it also has the disadvantage that the old PageRank value is higher than the new one.

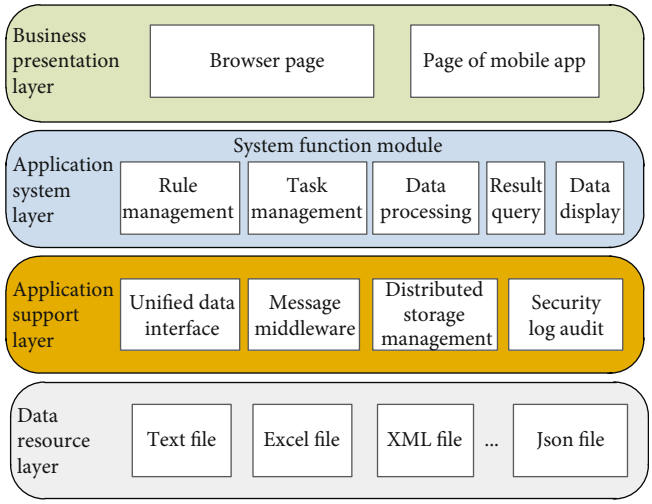

Figure 2: The system framework.

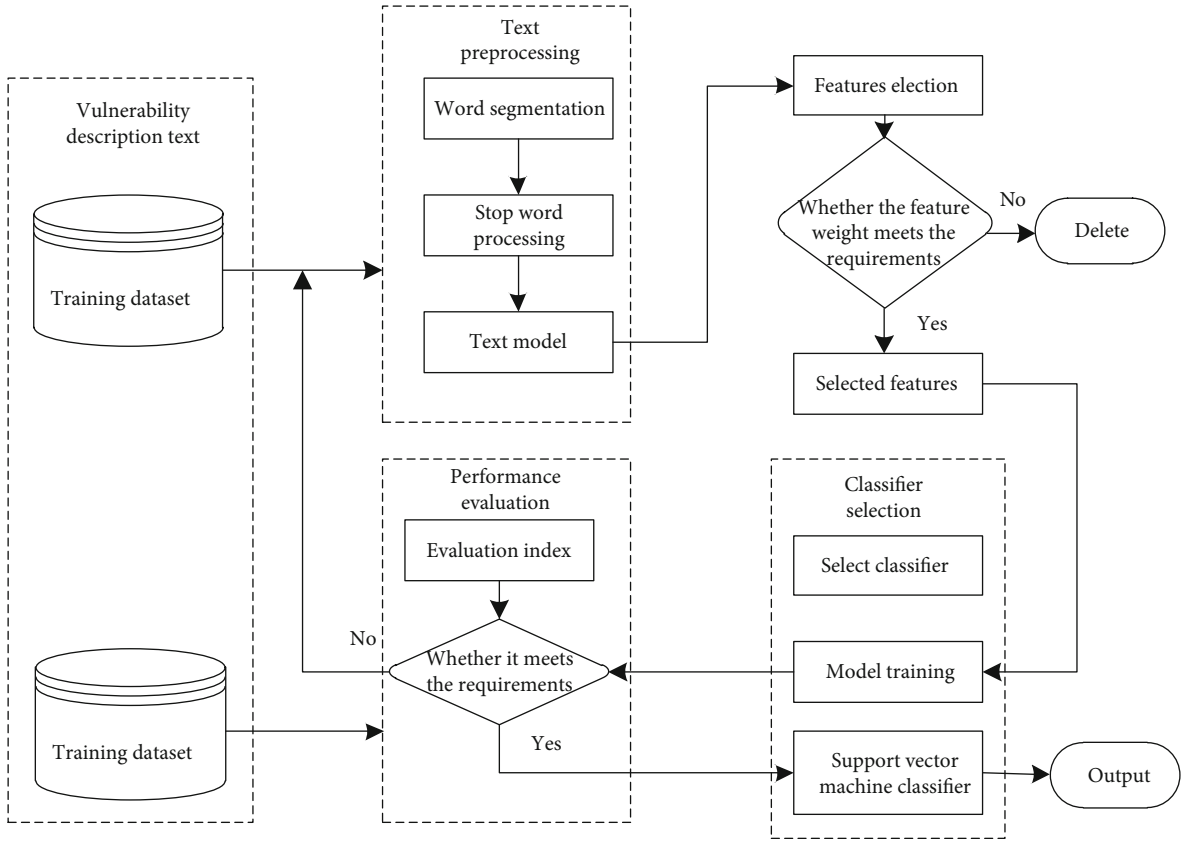

Figure 3: The process of text classification.

Because the keywords of text are similar to those of web pages, PageRank can be used to extract the keywords of text by modifying it to TextRank. The calculation of the TextRank value is as follows.

$$V(W_k) = (1-d) + d * \sum_{W_j \in I(W_k)} \frac{\text{link}(W_k, W_j)}{\sum_{W_t \in O(W_j)} \text{link}(W_t, W_j)} V(W_j). \tag{6}$$

In formula (6), $V(W_k)$ is the importance value of keywords, $\text{link}(W_k, W_j)$ is the connection between words, $I(W_k)$ is the word set of word W_k, $O(W_j)$ is the word set of word W_j, and the damping factor D is 0.85. For example, the description of CNNVD information security vulnerability no. is cnnvd-201905-408. "There is a vulnerability in Microsoft Windows with security features. The vulnerability is due to the lack of authentication, access control, rights management, and other security measures in the network

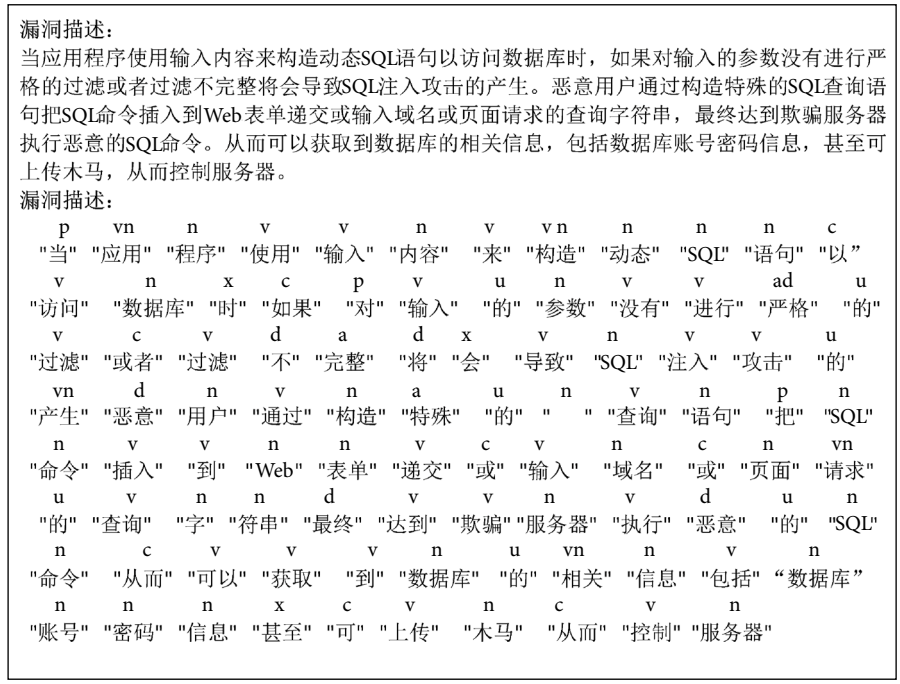

Figure 4: Examples of text segmentation.

system or product." After word segmentation, we get "Microsoft Windows / presence / security / feature / problem / vulnerability / network system / lack / authentication / access / control / authority / management / security measures." Then, we extract keywords and filter them according to the part of speech. The sliding window span is 5 and 7. The evaluation results are shown in Table 1.

In Table 1, precision refers to the actual positive data among all predicted positive data, which describes whether the prediction is accurate. Recall rate refers to the probability that all the real positive data are detected as positive examples, which represents the degree of integrity of all positive data. F1-measure combines the above two indexes of precision and recall, which represents the performance of classification. If the F1-measure value is higher, the performance of the classifier will be better.

5.2. Attribute Extraction Based on Frequency. Frequency-based attribute extraction is to identify and extract entity attribute information by making statistics of word frequency in vulnerability description text. Firstly, the middle noun of the comment sentence is identified by a POS tagger. The starting point of this idea is that the words with high frequency are important attribute words. Therefore, low-frequency words are usually not regarded as important words, and frequently occurring phrases are often important naming entities in this field.

The basic assumption of this method is that there are many text description information of vulnerability, and it is aimed at the same vulnerability, such as SQL injection. For example, a mutual information (PMI) score was used to calculate candidate phrases and entity classes with a "part whole relationship," in which the calculation formula of PMI is as follows.

$$\mathrm{PMI}(x, y) = \frac{h(a \cap d)}{h(a)h(d)}, \tag{7}$$

In the above formula, a is the candidate attribute word identified by the word frequency statistical method, D is the indicator word, and the search engine calculates the frequency information of word occurrence and cooccurrence. When the PMI value is too small, it means that a and D will not coexist frequently, which may not be a component.

For example, for CNNVD (China National Information Security Vulnerability Database) information security vulnerability description numbered cnnvd-201903-843, the word frequency is extracted. The extraction results of the first 16 words are shown in Table 2. The weight in the table is calculated based on the vulnerability description information and the existing records in the database.

Network management includes two aspects: network equipment management and network performance management. Network equipment management requires remote management and maintenance through the monitoring and parameter adjustment of the primary operation of the equipment to ensure the availability and safety of the network; network performance management ensures the reliability and efficiency of the network and optimizes the quality of the network through the monitoring and adjustment of various performance indicators. In order to achieve a unified and efficient management, it is necessary to conduct a comprehensive analysis of the problems in the whole system and analyze the network events together with the system,

Table 1: Assessment results.

Value of sliding window	Precision	Recall	F1-measure
Span = 5	0.7341	0.7173	0.7524
Span = 7	0.7412	0.7231	0.7572

Table 2: Word frequency analysis.

No.	Keywords	Word frequency	Normalized weight
1	Postscript	3	1
2	Artifex	3	1
3	Software	3	1
4	Ghostscript	2	0.9354
5	PostScript	2	0.9354
6	Loophole	2	0.8874
7	Open source	1	0.8471
8	United States	1	0.8314
9	Desktop	1	0.8302
10	Program	1	0.825
11	Attacker	1	0.824
12	Safety	1	0.8239
13	Parsing	1	0.8239
14	Table of contents	1	0.8231
15	Access	1	0.8201
16	Page	1	0.8181

database, and application events, so as to analyze the root causes of the problems. It can easily manage the switching network through the graphical interface, including remote monitoring, management and configuration of equipment, division and configuration of VLAN, and monitoring of network traffic.

6. Conclusion and Discussion

Firstly, this paper introduces the research status of information security vulnerability and machine learning identification domestically and internationally, including NVD in the United States, CNVD of the national information security vulnerability sharing platform in China, and detection system of network security vulnerability and the application of machine learning in network security, and then expounds the related concepts and technologies of information security vulnerability identification, including vulnerability types, text classification, and machine learning algorithm. Then, it analyzes the requirements of a vulnerability identification system, including the identification model, system requirements, and functional requirements, and introduces the design of a security vulnerability identification system, including the overall framework design, functional module design, database design, error tolerant security design, and text classification design based on the above design; it gives the information security vulnerability. The system implementation of the identification system and the key technologies of vulnerability text classification are introduced.

With the continuous advancement of network informatization, information security vulnerability identification will face greater challenges. The article has carried out some exploration and research on the security vulnerability identification system. Future research work can be improved from the following aspects. (1) The description of vulnerability text features can be improved. Since vulnerability text feature description is the basis of information security vulnerability identification, whether its representation is universal and accurate is critical to the security vulnerability identification system. Therefore, in-depth research will be conducted in the future on the characteristics of the vulnerability text so as to be able to filter out more accurate text feature items and improve the recognition accuracy and efficiency. (2) The design and implementation of the classifier can be improved. Feature selection is a key content in the text classification process, and the quality of the selection often directly determines the final classification result. Therefore, in addition to using common feature algorithms, some new algorithms and combination algorithms can also be tried, such as applying deep learning to the implementation of classifiers or combining multiple algorithms [26] so as to extract more accurately and efficiently and identify the characteristics of the vulnerability text.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the project of Shanghai Philosophy and Social Sciences Plan (No. 2018BGL023).

References

[1] Y. A. Basallo, V. E. Senti, and N. M. Sanchez, "Artificial intelligence techniques for information security risk assessment," *IEEE Latin America Transactions*, vol. 16, no. 3, pp. 897–901, 2018.

[2] J. F. Luo and D. N. Fu, "Analysis of computer virus epidemic situation in December 2018," *Netinfo Security*, no. 2, pp. 85–85, 2019.

[3] U. K. Singh, C. Joshi, and N. Gaud, "Information security assessment by quantifying risk level of network vulnerabilities," *International Journal of Computer Applications*, vol. 156, no. 2, pp. 37–44, 2016.

[4] N. Abu-Ghazaleh, D. Ponomarev, and D. Evtyushkin, "How the spectre and meltdown hacks really worked," *IEEE Spectrum*, vol. 56, no. 3, pp. 42–49, 2019.

[5] T. M. Conte, E. P. DeBenedictis, A. Mendelson, and D. Milojicic, "Rebooting computers to avoid meltdown and spectre," *Computer*, vol. 51, no. 4, pp. 74–77, 2018.

[6] A. Shahzad, S. Musa, M. Irfan, and S. Asadullah, "Key encryption method for SCADA security enhancement," *Journal of Applied Sciences*, vol. 14, no. 20, pp. 2498–2506, 2014.

[7] H. Yang, Y. Zhang, Y. Zhou, X. Fu, H. Liu, and A. V. Vasilakos, "Provably secure three-party authenticated key agreement protocol using smart cards," *Computer Networks*, vol. 58, no. 1, pp. 29–38, 2014.

[8] K. B. Shi, *Research on Information Security Leakage Identification System Based on Machine Learning*, Fudan university, 2018.

[9] G. Goth, "Functionality meets terminology to address network security vulnerabilities," *IEEE Distributed Systems Online*, vol. 7, no. 6, pp. 4–4, 2006.

[10] H. Holm, M. Ekstedt, and D. Andersson, "Empirical analysis of system-level vulnerability metrics through actual attacks," *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 6, pp. 825–837, 2012.

[11] C. Z. Cui, "Venustech's continuous construction of the information security ecological chain — analyze the information and cyber security strategy of Venustech," *Journal of Information Security Research*, vol. 3, no. 2, pp. 98–115, 2017.

[12] J. Diamant, "Resilient security architecture: a complementary approach to reducing vulnerabilities," *IEEE Security & Privacy Magazine*, vol. 9, no. 4, pp. 80–84, 2011.

[13] H. Yang and S. S. Lam, "Scalable verification of networks with packet transformers using atomic predicates," *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 2900–2915, 2017.

[14] Y.-H. Kim and W. H. Park, "A study on cyber threat prediction based on intrusion detection event for APT attack detection," *Multimedia Tools and Applications*, vol. 71, no. 2, pp. 685–698, 2014.

[15] J. Kim, Y. Zhou, S. Schiavon, P. Raftery, and G. Brager, "Personal comfort models: predicting individuals' thermal preference using occupant heating and cooling behavior and machine learning," *Building and Environment*, vol. 129, pp. 96–106, 2018.

[16] S. Martin, B. David, and H. Tracy, "Researcher bias: the use of machine learning in software defect prediction," *IEEE Transactions on Software Engineering*, vol. 40, no. 6, pp. 603–616, 2014.

[17] A. S. A. Aziz and A. E. Hassanien, "Multilayer machine learning-based intrusion detection system," in *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, vol. 70, pp. 225–247, 2014.

[18] X. He, S. Liu, and J. G. Jiang, "Comparative study of intrusion detection methods based on machine learning," *Netinfo Security*, vol. 18, no. 5, pp. 1–11, 2018.

[19] M. H. M. Hanif, K. S. Adewole, N. B. Anuar, and A. Kamsin, "Performance evaluation of machine learning algorithms for spam profile detection on Twitter using WEKA and RapidMiner," *Advanced Science Letters*, vol. 24, no. 2, pp. 1043–1046, 2018.

[20] P. Gogoi, D. K. Bhattacharyya, B. Borah, and J. K. Kalita, "MLH-IDS: a multi-level hybrid intrusion detection method," *The Computer Journal*, vol. 57, no. 4, pp. 602–623, 2014.

[21] W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system," *Expert Systems with Applications*, vol. 67, pp. 296–303, 2017.

[22] "China National Vulnerability Database of Information Security," *CNNVD vulnerability classification guide*, 2020, http://www.cnnvd.org.cn/web/wz/bzxqById.tag?id=3&mkid=3.

[23] L. Yang and S. Zhang, "A sparse extreme learning machine framework by continuous optimization algorithms and its application in pattern recognition," *Engineering Applications of Artificial Intelligence*, vol. 53, pp. 176–189, 2016.

[24] W. Liang, W. Huang, J. Long, K. Zhang, K. Li, and D. Zhang, "Deep reinforcement learning for resource protection and real-time detection in IoT environment," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6392–6401, 2020.

[25] P. Sunita and S. Jyoti, "A review of intrusion detection technique using various technique of machine learning and feature optimization technique," *International Journal of Computer Applications*, vol. 93, no. 14, pp. 43–47, 2014.

[26] W. Liang, Y. Fan, K.-C. Li, D. Zhang, and J.-L. Gaudiot, "Secure data storage and recovery in industrial blockchain network environments," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 10, pp. 1–6552, 2020.