ISSN: 2664-9640

Analysis of network information security protection strategy based on data mining algorithm

Guofeng Fu, Pengyi Li and Shiyan Gong

School of Yan'an University, Yan'an 716000, China

Abstract

With the continuous development of Internet technology and the continuous application of cloud computing technology, we have entered the era of big data. The emergence of data mining technology makes people's ability of data processing and analysis to a new level. This paper studies the process and main tasks of data mining, analyzes various technologies of network information security, puts forward the network information security strategy based on data mining, and improves the difficult problem of dealing with large amount of data in the network information security strategy.

Keywords

Data mining; Network information security; strategy.

1. Introduction

With the continuous popularization of current network applications, the requirements of government Internet Engineering and enterprise Internet engineering are constantly put forward. There are more and more challenges from network information security, and the requirements for network information security are also continuously improved. Most enterprises adopt more security products to strive to protect larger network systems and more information assets. However, the massive security events generated by each security product cause serious information overload. At the same time, the massive security events are full of a large number of unreliable information. The administrator has been helpless in front of these massive events. At the same time, all kinds of security products "go their own way", the security information between different products cannot be shared, forming "information islands", and the administrator lacks the ability to deal with mixed security threats. Therefore, there is a demand for centralized security management platform in the field of information security. Network centralized security management refers to the centralized and unified management of various security products. It centrally collects the security events, security vulnerabilities, log information, operation faults and other contents generated by various security products, and converts the data into intuitive, clear Actionable information is integrated into a unified security management platform to support the rapid response of the security system. Network centralized security management needs to deal with a large number of security events, so data mining technology is one of the core technologies. Through it, it can effectively support the association analysis of various information and mine valuable content from a large number of events.

2. Related concepts of data mining

2.1. Definition of data mining

Data mining is a method to extract uncertain and unknown information through relevant computer methods and algorithms in some irregular, heterogeneous and skilled huge data. The data source of data mining should be large and real, and the information we find should be useful and valuable to us. Theoretically, the larger and more random the amount of data, the

ISSN: 2664-9640

more accurate, representative and valuable the results obtained by data mining, which puts forward high requirements for the efficiency of relevant algorithms and technologies of data mining. Data mining is an interdisciplinary subject, which integrates the theories and technologies of database, artificial intelligence, statistics, machine learning and so on. Database, artificial intelligence and mathematical statistics provide three technical supports for the research of data mining. Data mining is a process that promotes some discrete, underlying and disordered large-scale data to orderly, acceptable and valuable knowledge by using relevant technical means, so as to provide help for decision-making. Specifically, data mining is to find out the internal laws and relationships between some data through the analysis of large-scale massive data. The specific process includes three stages: data preparation, information mining and result expression.

2.2. Main tasks of data mining

The main tasks of data mining include supervised learning, association analysis or frequent pattern analysis, clustering analysis, anomaly detection, etc.

Supervised learning includes two forms: classification and prediction. It refers to predicting new samples according to the size and type of known samples. Correlation analysis or frequent pattern analysis refers to finding such a regular connection pattern that when one event occurs, another event will also occur. Cluster analysis refers to finding out some internal laws and characteristics of all data, and dividing the data source into several data clusters according to these characteristics. Anomaly detection is to establish a template of data samples, and compare and analyze the data in the data source to find out the abnormal samples.

3. Basic theory of network information security

3.1. Definition of network information security

The concept of information security has not been put forward for a long time, and there is no unified understanding between countries. At the 56th UN General Assembly in 2001, the United Nations called on members of all countries to unify the basic concepts related to information security, so as to better deal with the issue of information security and strengthen international exchanges and cooperation. The understanding of network information security also has its development process. At present, scholars in various countries have a relatively consistent understanding that network information security can be divided into five characteristics, that is, according to the provisions of the Federal Information Security Management Act of 2002, network information security includes information integrity, confidentiality, availability, controllability and anti repudiation. Chinese scholars divide network information security into four aspects: environmental security, system security, program security and data security; There are also views that network information security includes database security, operating system security, virus prevention, network security, encryption and authentication, access control and so on. Network information security is a comprehensive subject involving network technology, computer science, cryptography, communication technology, applied mathematics, information security technology, information theory, number theory and other disciplines. In a broad sense, the relevant theories and technologies related to the integrity, confidentiality, authenticity, availability and controllability of information on the network are within the scope of network information security; In a narrow sense, network information security; It refers to the security of services and information in the network to ensure the security of software, hardware and system data in the network system.

ISSN: 2664-9640

3.2. Related technologies of network information security

3.2.1. Reptile Technology

Web crawler, also known as robot or spider, is a program that can automatically download web pages. There are tens of thousands of web pages on the Internet, which exist on various servers all over the world. Users can switch and browse each web page directly through web page links, and crawlers imitate human behavior, download or access multiple sites or web pages, and then hand them over to the data processing module.

3.2.2. Structured data extraction

Web Information Collection refers to the analysis of target information from a web page. It usually includes two problems. The first is to extract information from natural language texts, and the second is to extract information from structured data of web pages. We call the program for extracting this kind of data wrapper. There are three methods for wrapper: manual method, wrapper induction and automatic extraction.

3.2.3. Rule engine technology

Once the data is obtained, we need to process and analyze it. There are several common Python based rule engines. Pyke is a knowledge-based expert system, which adopts a language specification similar to Prolog. Prolog is a logic programming language, which is widely used in the field of artificial intelligence. Pychinko is a rule engine that can handle the semantic web, which can be defined by RDF. Intelligent is a rule engine based on domain specific language (DSL). It can define some rule expressions to monitor network data. A rule engine is an application that creates, stores, and manages rules, then executes them and infers other facts. The rules mainly refer to enterprise or business logic, legal terms, etc. In the development of rule engine, Rete algorithm and Prolog language are two important theoretical branches. Most rule engines are extended based on the above two. In industrial activity casting, clips system and Prolog system are two systems that have developed for a long time and are widely used.

4. Network security based on Data Mining

4.1. There are hidden dangers and vulnerabilities in the computer

On the one hand, there are defects in the computer itself, on the other hand, personal operation is not standardized, which makes the network environment in a hidden danger state, and it is difficult to ensure the security of the computer network. Therefore, to reduce the hidden dangers of computer network security, we not only need to continuously improve the security awareness and maintain the security of computer system, but also need to design or introduce more efficient security defense mechanisms to avoid hackers' attacks on computers, ensure the network environment and information security of enterprises, and improve the social competitiveness and internal environment of enterprises. Therefore, it is necessary to ensure the security of enterprise information and network resources and optimize the network environment protection of authorized users in the monitoring of hidden dangers and vulnerabilities of enterprise network computer security. However, at present, the security defense role of existing network authentication and firewall technology is limited, so it is difficult to avoid unauthorized users from taking advantage of computer network system vulnerabilities, Cause information leakage or network damage.

4.2. Network information tampering

At this stage, computer network is widely used in all walks of life. There is the shadow of computer network in the development of all fields. It is dependent and valued by government departments, institutions and economic enterprises. Although it is convenient to communicate and communicate through the network, there are also serious hidden dangers. Once the

ISSN: 2664-9640

computer network is stolen by hackers, the confidential documents and important databases stored in the computer will be stolen, causing large-scale security accidents and economic losses. Information leakage or theft events emerge one after another all over the world every year, because there are countless events caused by information security. These events and problems have gradually attracted people's attention, It has become an important potential hidden danger in international development, so various countries and enterprises are gradually strengthening computer network security management, and computer network security technology also needs further research and development.

4.3. There are still some defects in network security detection and evaluation technology

At present, in order to resist network security problems and protect computer data and information, network security detection and evaluation technologies summarize experience from actual use, test computer systems according to network attack experience, and realize the process from rule extraction to rule coordination. At present, a more mature computer security vulnerability detection technology - network security scanning technology, According to the rule evaluation method, the computer system is scanned, but there are still some defects. It can only detect the known computer network vulnerabilities, and it is difficult to evaluate the potential security risks. It is difficult to overcome the fact that rule generation depends on the empirical knowledge of component relationships, which improves the difficulty of rule generation. At this stage, network intrusion starts with the potential hidden dangers of computer system, and most of them inject viruses into the deficiencies in computer models. Therefore, in order to ensure computer security, the development of network security detection and evaluation technology is changing from the perspective of computer defects to the perspective of intruder intrusion, So as to turn the computer security vulnerability monitoring process into an anti intrusion process, but there are still some problems and deficiencies in the current development.

5. Construction of network information security model based on Data Mining Technology

5.1. Data mining algorithm in network information security management model

Data mining algorithms in network information security management model mainly include classification algorithm, sequence analysis algorithm and so on. On the one hand, the classification algorithm mainly constructs a perfect classification attribute model by using the form of analysis data training. Based on the construction of data training classification model, the attributes of the corresponding model can be described in a unified way. In the specific model attribute analysis link, most of them use the tuple analysis of relevant model database to analyze the attribute category of a data training sample. Through supervised data training, combined with the corresponding mathematical formula, the classification rules can be clarified step by step. Finally, after the classification rules of the corresponding model are clear, the prediction accuracy of the model can be analyzed one by one according to the corresponding classification rules. Combined with the setting of classification accuracy test samples, the test set model can be determined step by step. On the other hand, the sequence analysis algorithm is mainly based on the correlation analysis of different data records. In the application process of sequence analysis algorithm, transaction sequence pattern mining can be carried out for corresponding events, so as to obtain the minimum support frequent sequence that meets the user's requirements. On the basis of previous audit data association analysis, the sequence mode can be selected in combination with specific data association rules,

ISSN: 2664-9640

and finally the original data sequence function can be set. Aiming at the correlation between network attack and time variable, data mining based on sequence analysis algorithm can gradually clarify the relationship between network attack and time according to correlation analysis.

5.2. Construction of intrusion detection pattern based on Data Mining

Intrusion detection system based on data mining mainly includes data collection, data preprocessing, data decision-making, data algorithm and so on "。 Firstly, data acquisition and preprocessing mainly intercept the data in the program in the corresponding computer network module, and then convert it into ASCII network data packet combined with specific types of network data information. By further classifying and processing the corresponding network packets, network packets with various network connection forms can be formed. The corresponding network connection form also has its unique data source IP address, which provides an effective basis for data mining to a certain extent. It should be noted that in the process of Data Mining Intrusion Detection, it is necessary to organically integrate the user behavior data in the corresponding computer network and the current computer business data, and screen out the unnecessary data, so as to ensure the efficient progress of data mining. Secondly, after the data target is determined, according to the corresponding data state, combined with the implementation of data noise removal measures, the integrity and practicability of the overall data processing can be guaranteed. In the data mining link, the mining engine needs to be set according to the corresponding association rule database, sequence pattern analysis algorithm, system network algorithm, window clustering analysis algorithm and other algorithms. In the specific data audit program, it is necessary to set the intrusion data rules in the data mining link, and then determine the final data mining mode combined with the comparative analysis of the internal rules of the database. Finally, in the data decision-making link, the corresponding execution instructions are issued mainly according to the test data of the previous data mining algorithm. If the intrusion data line system is a malicious behavior, the corresponding computer system will automatically send out early warning information and take corresponding defense measures such as disconnection and port closure. On the contrary, it is allowed and continuously monitored.

5.3. Anomaly risk detection based on Data Mining

Network anomaly risk detection based on data mining mainly includes misuse detection and anomaly risk detection. Misuse detection is to complete the security detection of the corresponding network environment through the sample training set according to a special mode of sample supply.

Anomaly detection is a unified analysis of the corresponding Internet anomalies according to the setting of traffic behavior model. Network anomaly detection mainly uses association analysis algorithm to analyze the abnormal behavior in the corresponding computer network, so as to ensure the overall anomaly detection efficiency. The anomaly risk detection system based on data mining mainly constructs the three-time handshake connection mode based on TCP, the basic network intrusion data (see Figure 1). In the three-time handshake mode description link, combined with the relevant contents of Markov model and HMM, the specific model of TCP execution link can be carried out based on the analysis of server and client state transition frequency. At the same time, combined with the reasonable adjustment of TCP protocol identification, the volume of characteristic database is set to realize the real-time monitoring of network anomaly risk in data mining.

ISSN: 2664-9640

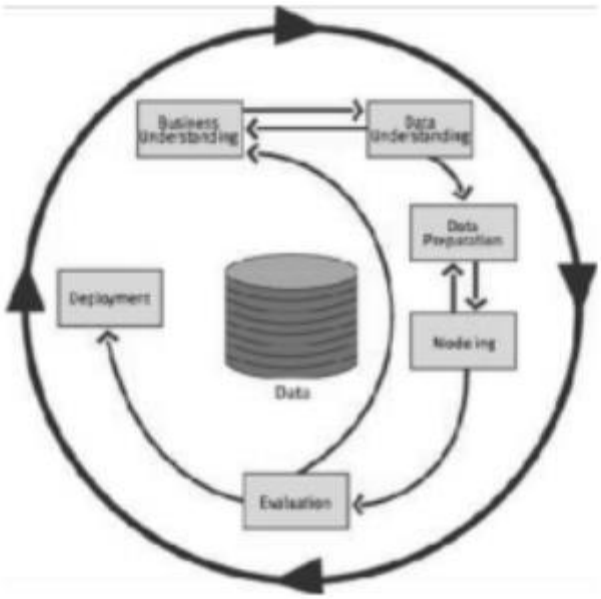

Figure 1 abnormal risk detection process based on Data Mining

6. Key points of network information security management supported by data mining technology

6.1. Building a secure network information environment

Safe network operation environment is the basis of data mining network information security management. In the actual network environment, it mainly includes the security of the overall computer network system, network intrusion detection, disaster recovery, network backup, anti-virus and so on. In the process of actual network environment security maintenance, isolated access control is mainly carried out for trusted and untrusted networks from physical and logical aspects. Through the management of user access network authorization, it can provide effective guarantee for basic network security management. On this basis, network intrusion detection technology can be adopted to uniformly analyze illegal intrusion, malicious damage and other behaviors. Combined with the establishment of malicious damage detection and early warning mechanism, security detection can be carried out within the corresponding computer network regularly to ensure the timely treatment of risk factors in the computer system. The operation of illegal intrusion detection and early warning mechanism can lay a solid foundation for the safe operation of the corresponding computer network operation environment. In the actual process of network information environment management, anti-virus technology can also be used to quantitatively analyze the overall network security threat. Through the centralized setting of virus protection, virus emergency, virus early warning and other systems, combined with the construction of audit analysis mode. The operation data of the computer system in the use link of the corresponding computer network can be analyzed uniformly, so that the malicious network attacks can be found in time according to the use of the system, so as to facilitate the smooth implementation of data mining. If the corresponding computer network system has been maliciously attacked, it is necessary to use the network backup and disaster recovery mode to recover the system within a certain time to maintain the stable operation of the overall network environment.

6.2. Ensure the security of data mining information

The security of data mining environment information is mainly based on the security of basic mining information storage, but also includes the security of data later transmission, use and so on. If the stiff wood creep detection and identification technology is adopted, it can effectively

ISSN: 2664-9640

identify the IRC / HTTP / P2P botnet control message, which mainly controls the botnet in the form of encrypted message. At the same time, the botnet is determined according to the IP address of the specified control end through behavior analysis. For the detection of suspected malicious code samples, the stiff wood creep detection and identification technology can detect the denial of service attack traffic through the traffic baseline, and export the parameters according to the specified time period, source, destination IP, protocol type and other data.

7. Summary

In the era of cloud computing and big data, computer network technology has been widely used in various industries of society. However, in the application of computer information technology, the frequency of illegal network intrusion and network malicious attack is also increasing. In this case, the traditional network security defense technology can not meet the needs of network information security management at this stage. Therefore, relevant personnel should pay attention to the effective combination of data mining technology and network information security management mode, so as to ensure the effective treatment of potential network threats and the safe operation of the overall computer network environment.

References

[1] Rybak I B , Wood D , Murray J , et al. SYSTEMS AND METHODS FOR EXTRACTION AND TELEMETRY OF VEHICLE OPERATIONAL DATA FROM AN INTERNAL AUTOMOTIVE NETWORK[J]. 2014.

[2] Il I . Clustering algorithms group objects into classes based on some measure of similarity (or distance) between objects, or an objective function that meas[J].

[3] Wang H , Xia H , Chen W , et al. Analysis of distributed dispatching automation system stability monitoring of topology based on divide-and-conquer strategy[J]. Power System Protection and Control, 2015.

[4] Yang Y , Wang C . A novel method of data correlation analysis of the big data based on network clustering algorithm[C]// IEEE International Conference on Communication Software & Networks. IEEE, 2015:360-366.

[5] Lai H , Lai X . Analysis and Application of Data Mining Based on Clustering Algorithm[C]// Information Technology & Mechatronics Engineering Conference. 2015.

[6] Hui Z . Analysis of International Marketing Strategy Based on Intelligent Mining Algorithm for Big Data[C]// 2018 11th International Conference on Intelligent Computation Technology and Automation (ICICTA). IEEE Computer Society, 2018.

[7]Shaozhen, Huang. Analysis of Students' Learning Behavior Based on Association Rule Mining Algorithm in Moodle Network Platform[C]// 2018.

[8] Shi L , Zhao H R , Zhang K . Research and Analysis of Network Data Mining Based on Genetic Algorithm[J]. Applied Mechanics & Materials, 2014, 651-653:2181-2184.

[9] Bursztein E , Mitchell J C . Using Strategy Objectives for Network Security Analysis[C]// International Conference on Information Security & Cryptology. Springer-Verlag, 2009.

[10] Liu B . Computer Network Information Security Protection Strategy Based on Clustering Algorithms[M]. 2020.

[11] Lin Z , Wei Z . Research on computer network information security and protection strategy[J]. Network Security Technology & Application, 2014.

[12] Zuo X , Chen Z , Dong L , et al. Power information network intrusion detection based on data mining algorithm[J]. The Journal of Supercomputing, 2020, 76(6–7).