Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 7895544, 10 pages
https://doi.org/10.1155/2022/7895544

Research Article

Visual Analysis of Multisource Heterogeneous Data Based on Improved DPCA Algorithm

Yun Zhou,[1] **Wen Mengfei**,[2] **He Youzhi**,[3] **Cheng Yun**,[3] **Lv Jinhui**,[1] **and Zuo Yi**[1]

[1]*Hunan University of Finance and Economics, Changsha, Hunan 410205, China*
[2]*Changsha College for Preschool Education, Changsha, Hunan 410007, China*
[3]*School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, China*

Correspondence should be addressed to Yun Zhou; zhouyun0110@vip.163.com

Received 8 March 2022; Revised 7 June 2022; Accepted 16 September 2022; Published 6 December 2022

Academic Editor: Yuxing Li

This paper proposes a multiview collaborative visual analysis system of network security based on a DPCA (clustering by fast search and find of density peaks) clustering algorithm called DPCANETVis, with network security analysis requirements for multisource heterogeneous data. Firstly, the system proposes an improved DPCA clustering algorithm based on the hierarchical relationship of mail sending and receiving to achieve the purpose of accurate classification. Secondly, a three-layer visual layout is designed to display relevant information such as data hierarchy, node relationship, behavior model, and other relevant information, by mixing a variety of interactive visual analysis methods such as tree diagram, word cloud, line diagram, subject river, and parallel coordinate. Lastly, based on the exploration of events, all suspicious nodes and their abnormal behaviors can be displayed in the system. Finally, the prototype system is used to analyze the network security log data set provided by the ChinaVis 2018, and the feasibility of the multilevel interactive visual analysis method for network security is verified through many experiments and discussions.

1. Introduction

With the increasing popularity of network communication technology and the rapid development of computer network, network office systems have become the mainstream resources and platforms in companies. However, there are also many network security problems, such as virus intrusion, hacker attack, and internal attack. In view of the above security risks, traditional network security technologies (including firewall, intrusion detection, host, and application status detection) [1–3] focus on the monitoring of abnormal events. And a large number of log files will be generated in the process, which lacks the ability of visual and real-time display and cannot achieve collaborative analysis among logs. This paper focuses on the abnormal analysis of the internal personnel behaviors in the enterprise information management system. These behaviors are legitimate from the perspective of the network, but combined with the context information such as the personnel department, the email sending and receiving, and account login, this behavior is a threat to the internal information security of the enterprise. Therefore, the traditional network security technology is not applicable to this kind of situation, and it is necessary to design a new detection system focusing on the internal personnel behavior of the enterprise.

Information visualization refers to the use of computer-supported, interactive, and visual representations of abstract data to enhance cognitive ability [4] and focuses on the presentation of implicit information and rules in data through visual graphics. Becker et al. [5] introduced the visual analysis technology into the network security log as early as 1995 and showed the network security state through data processing visually. Subsequently, the visualization field of the network security log was expanded, and the analysis technology was also improved gradually. Zhao et al. [6] classified and sorted the existing achievements from the perspectives of network security issues and network security visualization methods. Firstly, according to the timing

characteristics of network monitoring data, the advantages of the line chart, bar chart, and stack chart are emphasized. Then, according to the multidimensional characteristics of traffic monitoring data, the characteristics of the scatter diagram and parallel coordinate are elaborated in detail. Finally, for the correlation analysis of network security events, the radar map has strong graphic performance and interactivity. And the advantages of radar map in describing abnormal events have been further elaborated and improved in literature [7]. Zhang et al. [8] classified the graph techniques used in them into three categories: simple graph, conventional graph, and novel graph, by analyzing and comparing the characteristics of different existing network security log visualization analysis systems, and introduced them in detail. In recent years, the research on network security situational awareness have obtained more attentions [9, 10]. Literature [11] summarized the research status and existing problems of network security situation awareness, and then pointed out that the visualization of network situation would become one of the hotspots in the future.

Network security log belongs to multisource data, and for large companies, its internal staff is numerous, and the management structure is complex, so the internal network log data are large and the structure is different. Therefore, it is required to combine the advantages of various visualization technologies and carry out a reasonable layout. Different systems [12–15] often adopt several visualization methods according to their data characteristics and conduct a visual analysis of logs in the way of multigraph linkage. This paper needs to understand the network situation, analyze the enterprise structure and the daily behavior pattern, and mine abnormal events through the visual analysis of the internal network log of the enterprise. To solve this problem, we design an interactive visual analysis system DPCANETVis, which combines machine intelligence and human intelligence to help enterprises find abnormal behaviors that threaten their security and interests based on internal network log data.

2. System Design

2.1. System Design Process. The visual analysis system DPCANETVis includes three modules as shown in Figure 1, which are data storage and preprocessing module, data mining module, and visualization module. First of all, Python is used to conduct unified cleaning of multisource data sets, establish a unified and effective time format, process invalid values and missing values, and store them in a unified database to form the initial data source. Then, an improved DPCA clustering algorithm is proposed to analyze the node membership relationship and associate the node source IP (SIP) and ID. Finally, a three-layer visualization module is proposed to assist the network center to analyze the behavior patterns and abnormal events of network nodes layer by layer with an interactive interface.

Through the above design, the visual analysis system in this paper needs to mine the complex information hidden in the multisource network log data under the premise of unknown network node attributes and affiliations. Due to the wide variety of data, and considering the integrity and

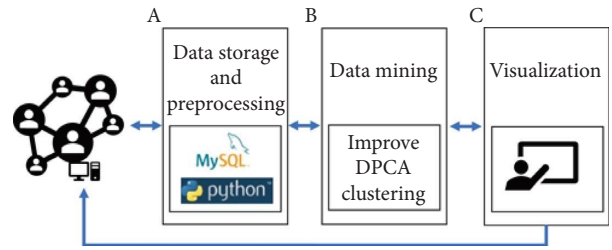


Figure 1: Visual analysis system architecture. DPCANETVis contains three modules: (a) data storage and preprocessing module; (b) data mining module; (c) visualization module.

logic of node behavior patterns and abnormal events, it is necessary to refine analysis objectives. After extensive conversations with the network hub staff, the following design goals are obtained:

R1: clustering network nodes accurately, and analyzing their attributes and hierarchical relations

R2: analyzing the behavior patterns of various nodes based on the e-mail semantics, and summarizing the semantic characteristics of emails to explore the behavioral differences

R3: assessing the abnormal risk by combining R1 and R2, and detecting the abnormal behavior of nodes in an interactive manner

2.2. Data Preprocessing

2.2.1. Data Cleaning. DPCANETVis is designed by using five types of monitoring data within an Internet company: login logs, web access logs, mail logs, TCPLOG, and clock in logs. As shown in Table 1, they mainly contain time, ID, application protocol, SIP, source port (SPORT), destination IP (DIP), and destination port (DPORT), etc. In detail, the web access log records the domain name information requested by the node; the mail log records the email addresses of the sender and receiver and the subject of the message; one or more TCPLOG records are generated by a node's login, web page visit, email sending or receiving, etc; each record gives the total number of bytes sent and received by the TCP connection; the time information in the punch card log contains the node's check-in and check-out time.

The data storage and preprocessing module mainly realizes the cleaning and filtering of the original data, which provides the basis for the subsequent data mining and analysis. This paper analyzes the mailbox composition of the internal nodes of the company through the mail log, filters the information, preserves the records that both send and receive meet the requirements, deletes the system mail and junk mail, and realizes the data filtering.

2.2.2. An Improved DPCA Algorithm. In order to obtain node attributes and their affiliations, this paper proposes an improved DPCA algorithm based on the mail sending-receiving relationship. The implementation of this algorithm is mainly divided into two parts. The first part is to carry out

Table 1: The format of data.

Form	Field name	Field meaning	Relevant description
Login.csv	Time	Log generation time	
	user	User name	Login user name
	proto	Applied protocol	SSH, mysql, and so on
	dip	Destination IP	Logged in IP
	dport	Destination port	Logged in port
	sip	Source IP	Login initiation IP
	Sport	Source port	Login initiation port
	State	Login results	Success or failure
Weblog.csv	Time	Log generation time	
	sip	Source IP	Client IP
	sport	Source port	Client application port
	dip	Destination IP	Server IP
	dport	Destination port	Server application port
	Host	Requested domain name	Host field of HTTP header
TcpLog.csv	stime	TCP data flow start time	The start time of the TCP stream, that is, the time when the first syn packet of the stream is received
	dtime	End time of TCP data flow	The end time of TCP flow, that is, the time when the last packet of the flow is received
	proto	Agreement	Protocol field value in IP packet header
	dip	Destination IP	Server IP of destination iptcp data stream
	dport	Destination port	Server application port of TCP data flow
	sip	Source IP	Client initiated IP of TCP data stream
	Sport	Source port	Client application port of TCP data stream
	uplink_length	Uplink bytes	The total number of bytes of application layer data sent from the client to the server from the establishment of the TCP stream to the end of the stream
	downlink_length	Downlink bytes	The total number of bytes of application layer data sent from the server to the client from the establishment of the TCP stream to the end of the stream
Email.csv	Time	Mail sending/ receiving time	Sending/receiving time of mail in the header
	proto	Application protocol	SMTP
	sip	Source IP	IP header source IP address
	Sport	Source port	TCP header source application port
	dip	Destination IP	IP header destination IP address
	dport	Destination port	TCP header destination application port
	from	Mail sender	From the corresponding field in the message header
	to	Mail recipient	It comes from the corresponding field in the mail header. When multiple recipients appear, they are separated by semicolons.
	Subject	Theme	From the corresponding field in the message header
Checking.csv	id	Employee ID	
	Day	Date	
	checkin	Check in time	
	checkout	Off duty sign off time	

word segmentation and destop word processing on the e-mail sending topics of nodes. Then, select a moderate number of high-frequency word as eigenvalues, and use the TF-IDF algorithm [16] to calculate the weight of the email topics of nodes under different eigenvalues to generate the high-dimensional coordinate matrix $T_{M*N} = \{x_1; x_2; \ldots; x_N\}$, where M is the total number of nodes, $n-1$ is the number of eigenvalues, and $N = 64$. In the second part, the node dependency structure is obtained through the following steps based on the hierarchical relationship between mail sending and receiving:

Step 1. Take the "summary" that appears in the subject of the message as the perspective to get the superior and subordinate relationship of the node: the mail receiver is superior, while the mail sender is the subordinate;

Step 2. Calculate T_{M*N} to generate a local density-distance coordinate system. The calculation formula of local density ρ_i is shown as follows:

$$\rho_i = \sum_{j \in I_s} e^{-\left(d_{ij}/d_c\right)^2}. \quad (1)$$

In the above formula, $I_S = \{1, 2, \ldots, M\}$, $d_{ij} = \text{dist}(x_i, x_j)$ which represents the Euclidean distance between vectors x_i and x_j, d_c is the truncation distance which is set to 0.5 in this paper. The calculation formula of distance a is shown as follows:

$$\delta_i = \begin{cases} \min\limits_{j<i} \left\{ d_{q_i q_j} \right\}, & i \geq 2, \\ \min\limits_{j \geq 2} \left\{ \delta_{q_j} \right\}, & i = 1, \end{cases} \tag{2}$$

where $j \in I_s$, $\{q_i\}_{i=1}^{M}$ is a descending order of $\{\rho_i\}_{i=1}^{M}$, and it satisfies $\rho_{q_1} \geq \rho_{q_2} \geq \cdots \geq \rho_{q_M}$. The decision graph as shown in Figure 2(a) is obtained by calculating the local density and distance.

Step3. γ_i considers the two indexes of local density and distance comprehensively, and the calculation formula is shown as follows:

$$\gamma_i = \rho_i \delta_i. \tag{3}$$

Then, the descending order of $\{\gamma_{ii=1}^{M}$ is obtained, as shown in Figure 2(b), and there is an obvious jump in the density of points above the red line: points above the red line (Figure 2(c)) are selected as the clustering center.

Step 4. For the other nonclustering center points, the classification attribute is given according to the nearest distance principle in the points whose local density is greater than itself.

Only leaf nodes are classified accurately according to the clustering results obtained through the above steps, which means upper-level nodes are classified to a same class, so it is difficult to obtain more layers of relationships. Therefore, the subordinate classification can be further improved by combining the superior and subordinate relationship obtained in Step 1.

Step 5. That is, if there is a hierarchical relationship between a node and a leaf node in the previous hierarchy, they will be grouped together.

Through the realization of the above two parts, the improved DPCA algorithm obtains the node dependency structure presented in the form of a tree graph, as shown in Figure 3. As can be seen from the figure, the leader in charge of the enterprise is 1067, and there are five department leaders, 1068, 1059, 1007, 1041, 1013, and 1068, respectively. However, the department type cannot be distinguished from the figure, and it is necessary to further mine the hidden rules through subsequent visual analysis techniques.

2.2.3. System Overview. DPCANETVis implements three different functions using a three-layer visual layout. As shown in Figure 4, the system consists of three pages: company organizational structure view, behavior pattern view, and risk assessment and analysis view. The view of company organization structure is based on tree diagram and word cloud technology, which explores the organizational structure and semantic information of network nodes and displays the email word clouds of different subordinate categories visually. The behavioral pattern view combines a multiline chart, stacked bar chart, and theme river chart to reflect the overall behavioral pattern and behavioral differences comprehensively based on data such as clocking log and server access. The risk assessment and analysis view is mainly based on a tree diagram, which combines parallel coordinate diagram, line diagram, and word cloud to carry out abnormal event monitoring, so as to further speculate the possible abnormal people and events through TCP traffic and mail word cloud.

3. Visual Analysis

3.1. Enterprise Organizational Structure View. The affiliation of nodes can be obtained by the improved DPCA algorithm, which belongs to a typical hierarchical structure. In order to further clarify various types of information and corresponding mail characteristics, the system adopts an interactive design of a tree diagram and a two-level word cloud.

As shown in Figure 5, the view is composed of (a), (b), and (c). Among them, (a) based on the radar tree diagram can clearly display node attributes and hierarchical relationships (R1). Radar tree diagram is a kind of tree diagram, which is more suitable for scenes where the depth of each branch is relatively consistent. Through the radar tree diagram, the distribution of each department and its number is understood, and a basic understanding of the organizational structure is formed.

By clicking on a node in the radar tree diagram, (b) will display the distribution of the mail subject of the node in the form of a word cloud, and (c) will also display the mail subject distribution of all child nodes (subordinates) of the node. The word cloud is used to highlight high-frequency keywords in emails and filter out a large amount of low-frequency information so that users can grasp the subject at a glance. According to the distribution of the subject of the mail, it can be inferred which category it is (R1). After that, the radar tree chart is color-coded to distinguish different departments of the enterprise.

3.2. Risk Assessment Analysis View. In order to analyze the behavior patterns of different types of nodes, this paper designs a behavior pattern view (R2) based on the check-in and network traffic information in the network log. As shown in Figure 6, first, the punch-card distribution in different periods of time is displayed with two line graphs (a) and (b), where each department is distinguished by different colors. The line graph is easy to show the changes in things with variables, and the punching rules of various nodes can be observed. Then, (c) displays the number of punch cards in each department in a month with a stacked histogram. As an extension of the histogram, the stacked bar chart superimposes each category so that it can display the total amount of each category and the size and proportion of each subcategory contained in the category. Therefore, through Figure (c), we can not only know the number of check-in nodes per day but also know the proportion of different types of nodes in it. In particular, by paying attention to the number of nodes in individual time periods, the special behavior patterns of different types of nodes can be known. Finally, (d) shows the daily flow usage of various nodes in the form of a thematic river graph. The theme river map is a

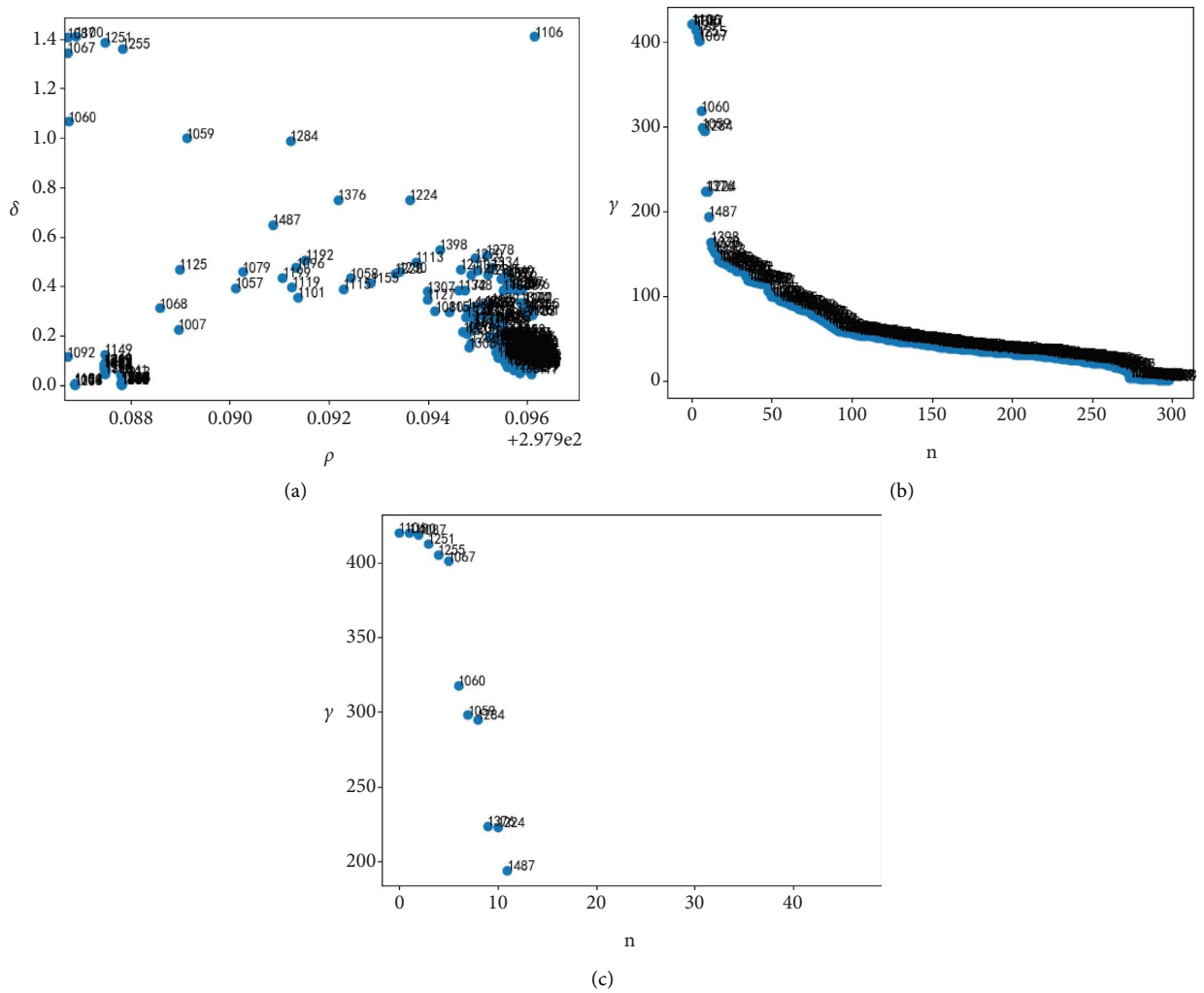


Figure 2: $\rho - \delta$ coordinate system point. (a) The decision graph. (b) The descending order diagram. (c) The clustering center point.

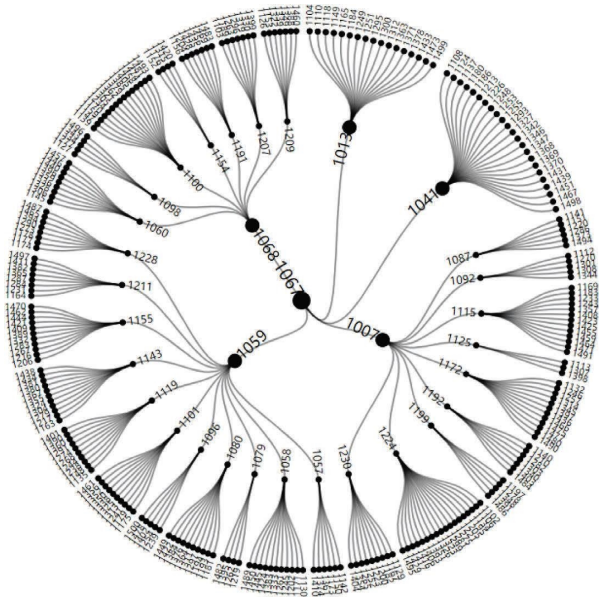
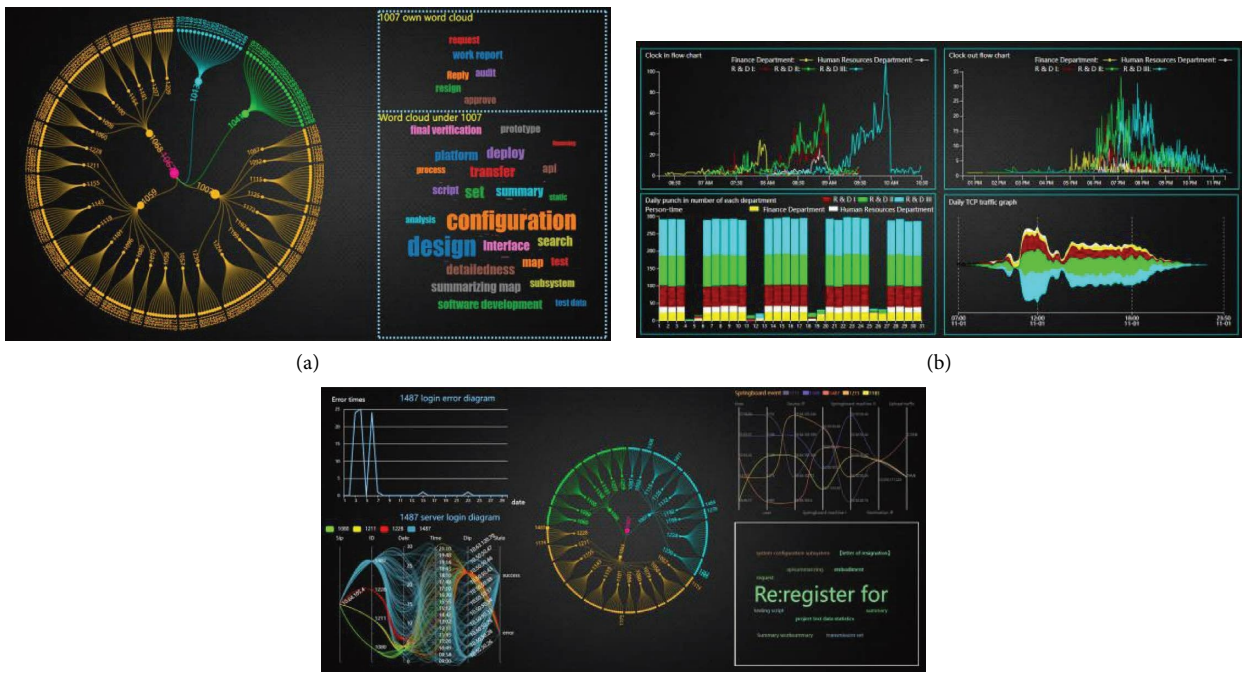


Figure 3: Node-dependent structure tree.



(a) (b)

(c)

Figure 4: DPCANETVis. (a) Company organizational structure view. (b) Behavior pattern view. (c) Risk assessment analysis view.

variation of the stacked area map, which expresses the changes of different types of data over time through the shape of the flow. In the figure, the peak period of network traffic is found, and the behavior of the node is represented.

3.3. Risk Assessment Analysis View. In visual design, word cloud images generally display the size of words according to word frequency from large to small. Based on node behavior patterns, this article assumes that the occurrence of low-frequency topics in emails may prompt dangerous information. In this way, an event-driven risk assessment analysis view is designed (R3). The article defines network security events as four elements with a logical sequence: mail receiving and sending, login errors, server login information, and springboard events. The specific design is shown in Figure 7.

Taking the middle radar tree diagram as the starting point, firstly, the target node is selected, the e-mail word cloud of this node is observed, and suspicious emails are found. Determine whether the node is suspicious by analyzing the number of incorrect logins and the subject of the mailbox. Then, from the log-in information parallel coordinate graph, the detailed log-in source IP, user number, time, destination IP, and whether it is successful can be observed. By analyzing the reason and time of the login error, situations such as account theft can be found. Finally, a parallel coordinate diagram of the springboard machine is designed to show the use of multihop servers in nodes. Based on this diagram, information such as source nodes, intermediate paths, destination nodes, and single upload traffic can be analyzed. The above event elements constitute a complete

Figure 5: The company organizational structure view. (a) Radar tree diagram of organizational structure. (b) Mail subject word cloud of the parent node. (c) Mail subject word cloud of the child node.

Figure 6: The behavior pattern view. (a) Line diagram of node clocking in the situation; (b) line diagram of node clocking out situation; (c) stacking bar chart of clocking of the total number; (d) theme river chart for flow.

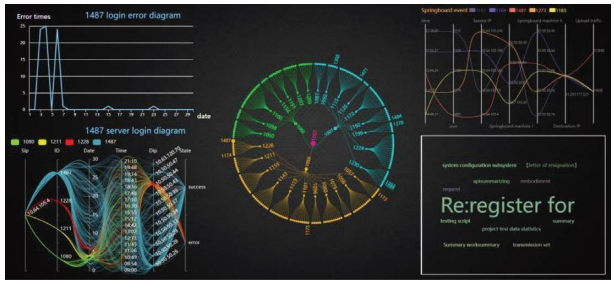


Figure 7: The risk assessment analysis view. (a) Improved word cloud (low-frequency displayed firstly); (b) line chart of logon error count; (c) parallel coordinate chart of server logon; (d) parallel coordinate chart of board divulge.

network abnormal event, which may include illegal uploading of internal data. Behaviors such as leaks can be analyzed.

4. Case Study

The data set [17] of ChinaVis 2018 Data Visual Analysis Challenge 1 is used to verify the effectiveness of the system in this paper. The background of the competition is that a high-tech company is preparing to release a new heavyweight product. In order to ensure the smooth release of the product, a visual analysis system is needed to analyze the recent internal work patterns of the company and assist intelligence personnel in discovering abnormal events. This data set contains one month's internal monitoring data of the enterprise, including login logs, web access logs, mail logs, check-in logs, and TCP traffic logs.

4.1. Analysis of Enterprise Organizational Structure. The intelligence personnel analyzes the enterprise organizational structure view, as shown in Figure 8. From the radar tree diagram, the enterprise hierarchy is roughly divided into three levels: general manager, department manager, and ordinary employee. Click the 1041 node (employee) in the tree, as shown in Figure 6. The word cloud linkage on the right shows the subject of the employee's emails sent and received. It is found that the main words are "finance" and "reimbursement," and it is analyzed that the employee belongs to the finance department, and its subordinate nodes are all employees of the finance department.

Further analysis revealed that the company is composed of 5 departments including the Finance Department, Human Resources Department, R&D Department 1, R&D Department 2, and R&D Department 3. Except that the R&D department has an additional team leader, which is composed of four levels, the other departments are composed of three levels, as shown in Figure 8.

4.2. Analysis of Working Mode. The behavioral pattern view is analyzed. Observe the daily commuting time of each department, and view the commuting curve of R&D 2 departments through interactive operation, as shown in Figure 9. It is speculated from the peak of the curve that the department's working time is 10 am and the end time is 8 pm, but most of them leave before 11 pm, and there is a common situation of working overtime at night.

Then, by looking at the theme river map and analyzing the flow access of each department, we can get the working mode: as shown in Figure 10, each department has a strong work intensity from 10 am to 12 o'clock in the morning, and the work is relatively stable in the afternoon. By comparing the flow of each department, it can be known that the three R&D departments have the most visits, which is not unrelated to the need to pay close attention to the latest technology;

Finally, the stacked histogram of the number of punches is analyzed. As shown in the left chart of Figure 11, it was found that the number of clock-in numbers in each department has little change in working days, so absenteeism rarely occurs. Put the time on November 19, 25, and 26, 2017. This day is a weekend, but almost all the staff in the finance department work overtime, as shown in the right figure of Figure 11. It is speculated that the company has heavy financial affairs at the end of the month, which may be a common case.

5. Enterprise Risk Assessment Analysis

Intelligence personnel interactively use risk assessment views to try to spot anomalies. According to the definition of event-driven, the analysis can be started from the tree diagram, the employees who log in incorrectly can be located from the suspicious e-mail word cloud, and the suspicious events can be analyzed. This view is used to get the following complete event analysis.

5.1. Three Persons Resign on the Same Day. Combined with the word cloud, we can interactively analyze any employee node in the radar tree diagram and found that employees 1281, 1376, and 1487 are all from the R&D department. They submitted their resignations on the same day and were approved by two department managers. It is speculated that there is a suspicion of abnormality here, and further analysis is required. As shown in Figure 12, the word cloud display of 1281 is found in the tree diagram.

5.2. No. 1487 Employee Embezzled the Group Leader Account Incident. As shown in Figure 13 on the left, it is a line chart of the number of login errors. It was discovered that employee No. 1487 had more than 20 login errors on November 3 and November 6. This further verifies the employee's anomaly. Observe the parallel coordinate diagram of the server login. As shown in the right figure of Figure 13, the employee's IP address tried multiple times to log in to the 1080 and 1211 group leader accounts on November 3 and 4 but failed after multiple attempts. On November 6th, he tried to log in to the 1228 group leader account several times. The login was successful at 22:00, and then, he logged in on the 16th and 24th. Therefore, it is speculated that the employee is suspected of stealing the team leader's account and leaking secrets.

Figure 8: The analysis of company organizational structure.

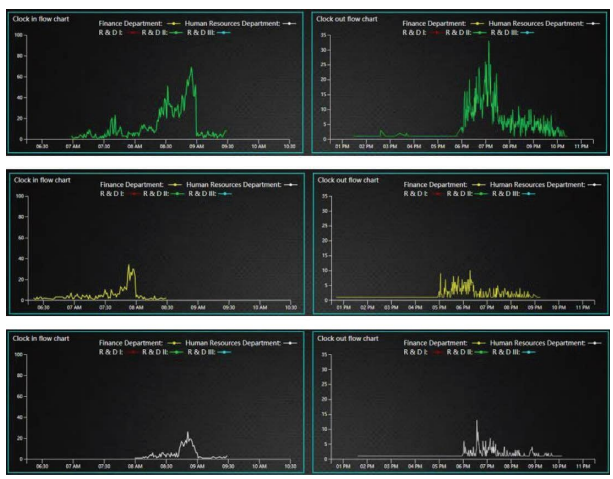


Figure 9: The clocking time of each department.

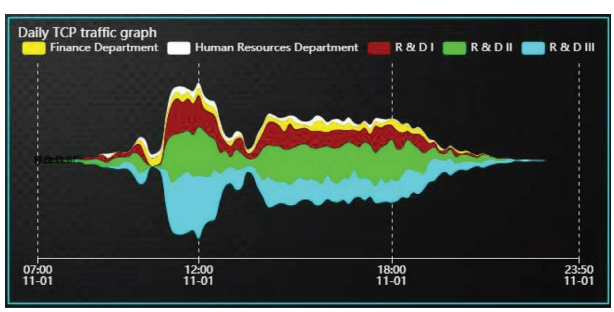


Figure 10: The flow usage of each department.

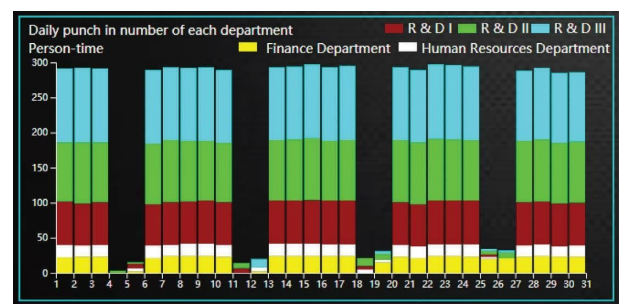


Figure 11: The number of clocking of each department.

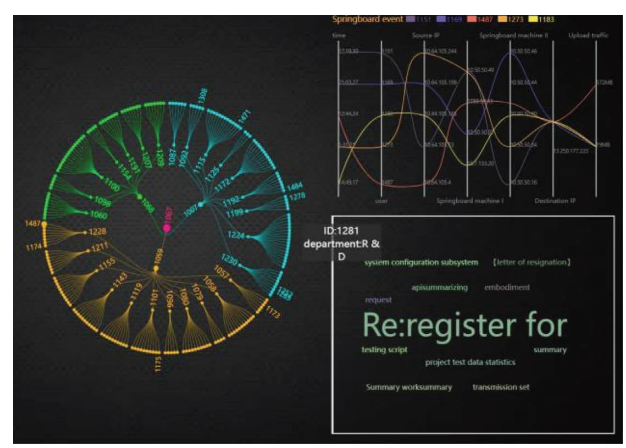


Figure 12: The low-frequency word cloud of 1281.

5.3. Group Leaks. In order to gain insight into the specific behavior of employee No. 1487 in embezzling the team leader's account, intelligence personnel used the springboard parallel coordinate map, as shown in Figure 14, and found that on the 24th, they used the team leader's account to upload 572 MB of data to overseas servers. During the analysis of 1487, it was also discovered that four other employees within the company had uploaded data to the server together. Through the word cloud, there are "resignation" and "recruitment" messages in the subject of the emails of several other employees, and these people were absent from work on the same day. Intelligence personnel believes that this is most likely a gang leak.

6. User Feedback

To verify the effectiveness of this system, we invited an enterprise manager and a network security expert to make a preliminary evaluation of the visual system. The former has clear requirements for security analysis, while the latter has extensive work experience in the field of network security.

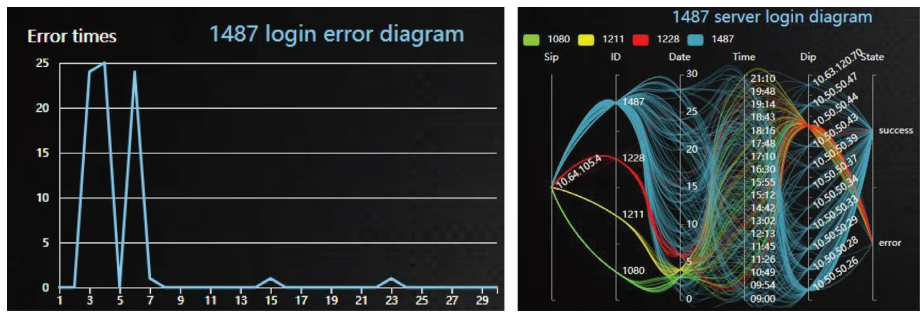


Figure 13: The logon error condition of 1487.

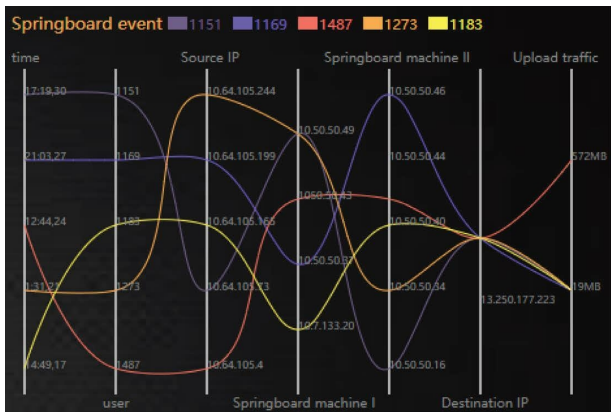


Figure 14: The springboard diagram of gangs leak.

Firstly, we briefly introduced them to the use process of the system and then collected feedback from experts after using the system. Combining expert feedback, the advantages and disadvantages of the visual analysis system proposed in this article will be discussed below.

The main advantages of the system are as follows: (1) based on the improved DPCA clustering algorithm, there is no need to set the number of clusters in advance, automatically analyze the internal organizational structure of the enterprise, and can effectively respond to changes in the network structure. (2) At the same time, frequent user settings are avoided, and the user's workload is reduced; visual technology is used to show the analysis process and improve work efficiency. (3) The risk assessment view allows users to participate in the abnormal discovery and analysis process in an interactive manner, which improves the interpretability of the analysis process.

By comparing with the answers published by the competition committee, it is proved that the analysis results of the system basically meet the standard answers.

The disadvantages are as follows: (1) although the DPCA clustering algorithm does not need to specify the number of clusters, it still requires users to set some parameters; (2) the system's ability to process larger-scale data needs to be verified. Web log data are a kind of streaming data, and as the scale of the enterprise increases, the amount of data will increase exponentially. How to improve the clustering algorithm and visual view to adapt to the increase in the amount of data is the main challenge.

7. Conclusion

This paper proposes an interactive visual analysis framework for multisource heterogeneous network log data. Compared with previous methods, this method has the following three advantages:

(1) By introducing interactive word cloud technology, the improved DPCA clustering included can be used to effectively dig out the node organization structure

(2) The method of multigraph dynamic linkage makes it easy to use and quickly master the organizational structure and working mode of the enterprise from the aspects of personnel structure, mail receiving and sending, clock in, and so on

(3) From the perspective of users, the system provides interactive visual analysis to effectively mine the insiders of enterprises stealing important data

In the future, research work will be carried out in the following areas: (1) isolated abnormal events need to be further analyzed. Some isolated events in the web logs cannot be specifically evaluated and dealt with because no relevant contextual information can be found, so further analysis is needed. (2) Case studies need to be enriched. Based on a single case may not be able to show the overall picture of the system, a larger data set will be used to fully demonstrate the research work of this article. (3) Scalability needs to be improved. The framework of this article is limited to specific network security log data and will combine more visualization techniques to improve the generalization of the system.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by theNational Natural Science Foundation of China (Grant no. 61902117) and National Natural Science Foundation of Hunan Province (Grant no. 2020JJ5010).

References

[1] F. Iglesias and T. Zseby, "Analysis of network traffic features for anomaly detection," *Machine Learning*, vol. 101, no. 1-3, pp. 59–84, 2015.

[2] L. I. Hong-lin, *Design and Implementation of Firewall Based on Decision Tree*, DLMU(Daliann Maritime University), Dalian, China, 2018.

[3] W. Xiong, H. P. Hu, N. X. Xiong et al., "Anomaly secure detection methods by analyzing dynamic characteristics of the network traffic in cloud communications," *Information Sciences*, vol. 258, no. 10, pp. 403–415, 2014.

[4] K. Moreland, "A survey of visualization pipelines," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 3, pp. 367–378, 2013.

[5] R. A. Becker, S. G. Eick, and A. R. Wilks, "Visualizing network data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 1, no. 1, pp. 16–28, 1995.

[6] Y. Zhao, X.-p. Fan, F.-f. Zhou, F. Wang, and J.-w. Zhang, "ASurvey on network security data visualization," *Journal of Computer Aided Design and Graphics*, vol. 26, no. 05, pp. 687–697, 2014.

[7] X. Fan, W. J. Luo, X. J. Dong, and R. Su, "A network visualization system for anomaly detection and attack tracing," in *Proceedings of the 2018 International Conference of Pioneering Computer Scientists, Engineers and Educators (ICPCSEE2018)*, pp. 1–15, Zhengzhou, China, September 2018.

[8] S. Zhang, J. Zhao, and R. Chen, "Research advances on network security logs visualization," *Journal of Frontiers of Computer Science and Technology*, vol. 12, no. 5, pp. 681–696, 2018.

[9] T. Bass and D. Gruber, "A glimpse into the future of id," *The Magazine of USENIX & SAGE*, vol. 24, no. 3, pp. 40–49, 1999.

[10] HSARPA Fact Sheet, Homeland Security, 2022, https://www.dhs.gov/publication/fact-sheet-hsarpa.

[11] J. Gong, X. D. Zang, Q. Su, X. Y. Hu, and J. Xu, "Survey of network security situation awareness," *Ruan Jian Xue Bao/Journal of Software*, vol. 28, no. 4, pp. 1010–1026, 2017.

[12] Y. Zhao, X. Fan, F. Zhou, W. Huang, and M. Tang, "Research on collaborative visual analysis of large scale network security Data," *Journal of Frontiers of Computer Science and Technology*, vol. 8, no. 7, pp. 848–857, 2014.

[13] H.-y. Jiang, Y.-d. Wu, S. U. Meng-xin, X. Wang, and Z. H. Yu-wei, "Research on fusion and visual analytic method of multi-source network security data logs," *Journal of Southwest University of Science and Technology*, vol. 32, no. 01, pp. 70–77, 2017.

[14] Y. Zhao, X. Liang, X. P. Fan, Y. W. Wang, M. J. Yang, and F. F. Zhou, "MVSec: multi-perspective and deductive visual analytics on heterogeneous network security data," *Journal of Visualization*, vol. 17, no. 3, pp. 181–196, 2014.

[15] I. Kotenko and E. Novikova, "Vissecanalyzer: A visual analytics tool for network security assessment," in *Proceedings of the International Conference on Availability, Reliability, and Security*, pp. 345–360, Regensburg, Germany, September 2013.

[16] C. H. Huang, J. Yin, and F. Hou, "A text similarity measurement combining word semantic information with TF-IDF method," *Chinese Journal of Computers*, vol. 34, no. 5, pp. 856–864, 2011.

[17] ChinaVis 2018, ChinaVis, 2018, https://chinavis.org/2018/index.html.