Research on Knowledge Graph Model for Cybersecurity Logs Based on Ontology and Classified Protection

Yuan Tao\textsuperscript{1,2}, Moyan Li\textsuperscript{1,2} and Wei Hu\textsuperscript{1,2}

\textsuperscript{1}The Third Research Institute of Ministry of Public Security, Shanghai, China
\textsuperscript{2}National Engineering Laboratory for Key Technology of Classified Information Security Protection, Beijing, China
Email: programfox@126.com

Abstract. In order to audit and analyse cybersecurity events from massive logs, the knowledge graph model for cybersecurity logs is proposed. The data layer and the pattern layer of the knowledge graph model for cybersecurity logs is obtained based on ontology and classified protection. Entity classification of cybersecurity logs is extracted by using classified protection standards in the data layer of the knowledge graph. The relationship and attribute of ontology from cybersecurity logs is defined. The pattern layer of the knowledge graph model for cybersecurity logs is defined, classified protection data is integrated in entity alignment, and classified protection information gain is used in ontology construction. The structure of the knowledge graph for cybersecurity logs is given. So that the efficient association and deep mining analysis of cybersecurity logs are realized, and it can be directly analysed and processed on the data without the need for precise modelling of the problem. The efficiency of logs analysis is improved. The model has heuristic characteristics and generalization abstract ability, and the model is suitable for big data analysis of large-scale cybersecurity logs.

1. Introduction

Important information system is an important basic of the country, which plays an important role in the normal information society. At the same time, because of the high value of information, the important information system has become the main target of hackers. Lots of security protection devices are used to protect important information system, and the massive cybersecurity logs are recorded.

In order to audit and analyse the cyberspace security events from these massive logs and trace the origin of events, the knowledge graph \cite{1-5} model for cyberspace security log based on ontology is built, so as to analyse and mine the association of massive multi-source cybersecurity logs.

The knowledge graph model for cyberspace security log is a network knowledge base on attribute entities linked by relationship. From the perspective of graph, knowledge graph is essentially a conceptual network, in which the nodes represent the entities of the physical information system, and the data flow relationships among entities constitute the edges of the network. The classified protection \cite{6-9} data is analysis fusion to gain data.

The logical structure of knowledge graph model for cyberspace security log is divided into two layers: the data layer and the pattern layer. The ontology is used in the pattern layer to describe the concept level system, so as to give a clear definition of cybersecurity logs and their relationships in a formal way. In this way, the cybersecurity logs can be shared and related to audit and analysis in different times and places \cite{10-12}.

The pattern layer is the core of the knowledge graph. In the pattern layer, the refined knowledge is stored. Ontology model is built to express the knowledge of cybersecurity and attack behaviours. In
this way, ontology can be used as a common language to express the knowledge of attack behaviours, so that the association of massive cybersecurity logs can be constructed and analysed [13-16].

2. Data Layer of the Knowledge Graph Model for Cybersecurity Logs

2.1. Entity Extraction of Cybersecurity Logs based on Classified Protection

Entity extraction of cybersecurity logs based on classified protection is automated to classify the massive cybersecurity logs from multi-source heterogeneity security devices. The quality of entity extraction has a great impact on the efficiency and quality of subsequent knowledge acquisition.

In order to get entity extraction, cybersecurity logs are classified according to the technical categories in “GB/T 22239-2019 Information Security Technology- Baseline for Classified Protection of Cybersecurity”, “GB/T 28448-2019 Information Security Technology- Evaluation Requirement for Classified Protection of Cybersecurity” and “GB/T 25070-2019 Information Security Technology-Technical Requirement of Security Design for Classified Protection of Cybersecurity”, so the entity classification of cybersecurity logs is as table 1 shown.

In table 1, Security management centre is a platform or area for unified management of security strategy and security computing environment of designated the information system. Security computing environment refers to the relevant components that store, process and implement security policies of the information system. Security area boundary refers to the boundary of security computing environment for the information system, the relevant components between security computing environment and security communication network, so as to realize connection and implement security strategy. Security communication network refers to the relevant components for information transmission and implementation of security strategy between security computing environments of the information system.

| Entity classification | Cybersecurity logs |
|-----------------------|--------------------|
| Logs of security management centre | (1) Logs of security devices, such as firewall, IDS, etc.  
(2) Logs of network devices, such as switches, routers, etc.  
Logs of security computing environment | (1) Logs of database, such as MySQL, ElasticSearch, etc.  
(2) Logs of operating system, such as Windows operating system, Linux operating system, and other UNIX operating system.  
(3) Logs of Cloud platform, such as VMware, OpenStack, etc.  
(4) Logs of big data platform, such as Hadoop, spark, storm, etc.  
(5) Logs of web application, such as Apache, IIS, Nginx, etc.  
(6) Logs of middleware, such as Tomcat, WebSphere, JBoss, etc.  
(7) Logs of Other application, such as bind, FTP, SSH, etc.  
Logs of security area boundary | (1) Logs of the remote desktop  
(2) Logs of SSH access  
(3) Logs of SMTP / POP3 / IMAP  
Logs of security communication network | (1) Logs of network connection, such as TCP / UDP / ICMP connection.  
(2) Logs of DNS  
(3) Logs of FTP  
(4) Logs of Http request and response. |

2.2. Relationship and Attribute Extraction of Cybersecurity Logs

The relationship between the entities is need to extract from the security attribute of the information systems, and the entities is linked through the relationship to form a network of knowledge structure. The relation extraction is to solve the problem, which is how to extract the relation between entities from cybersecurity logs.
Relationship is the basis of building ontology. The quality of relationship is directly determined the quality of the ontology and the whole knowledge graph. In order to extract relationship from entity classification of cybersecurity logs, the attributes of important information systems are needed to classify. So that the security attributes of important information systems are needed to be further extracted and analysed.

The ontology relationship of cybersecurity logs is described from four aspects: system type, system applicability, system confidentiality and system integrity. The concepts and assign values of the relationship and attribute of ontology from cybersecurity logs is shown in Figure 1.

| Relationship and attribute of ontology from cybersecurity logs |
|---------------------------------------------------------------|
| **System type (T)**                                          |
| - **Classified level (T₁)**                                  |
| - **Technology type (T₂)**                                   |
| - **Service type (T₃)**                                      |
| **System applicability (A)**                                 |
| - **Usage frequency (A₁)**                                   |
| - **Recovery time requirements (A₂)**                        |
| - **Unavailable loss size (A₃)**                             |
| **System confidentiality (C)**                               |
| - **Data type (C₁)**                                         |
| - **Data leakage loss (C₂)**                                 |
| - **Social impact of data leakage (C₃)**                      |
| **System integrity (I)**                                     |
| - **Integrity check required (I₁)**                          |
| - **Scope of impact of integrity damage (I₂)**               |
| - **Loss of integrity damaged (I₃)**                         |

**Figure 1.** The relationship and attribute of ontology from cybersecurity logs

The goal of attribute extraction is to collect attribute information of cybersecurity logs from different variety of log sources. The attribute of system type is described from three aspects: classified level (T₁), technology type (T₂) and service type (T₃). The optional values of each attribute are: T₁ = {the first level, the second level, the third level, the fourth level, the fifth level}, T₂ = {general technology, cloud computing technology, Internet of things technology, mobile Internet technology, big data technology, industrial control technology}, T₃ = {town, city, province, country}.

The attribute of system applicability is described from three aspects: usage frequency (A₁), recovery time requirements (A₂) and unavailable loss size (A₃). The optional values of each attribute are: A₁ = {low, commonly, frequently, very frequently}, A₂ = {long, general, short, real time}, A₃ = {indirect, small, medium, large}.

The attribute of system confidentiality is described from three aspects: data type (C₁), data leakage loss (C₂) and social impact of data leakage (C₃). The optional values of each attribute are: C₁ = {public, internal, sensitive, high sensitivity}, C₂ = {indirect, small, medium, large}, C₃ = {indirect, small, medium, large}.

The attribute of system integrity is described from three aspects: integrity check required (I₁), scope of impact of integrity damage (I₂) and loss of integrity damaged (I₃). The optional values of each attribute are: I₁ = {require, no-require}, I₂ = {internal, external, both internal and external}, I₃ = {small, medium, large}.

In relationship and attribute extraction of cybersecurity logs, relationship can connect cybersecurity logs with important information system, and achieve a certain role of knowledge connection. The entity classification based on classified protection of cybersecurity is mapped to information system and cybersecurity logs, such as domain, classified protection data, range, and so on.

3. Pattern Layer of the Knowledge Graph Model for Cybersecurity Logs

The pattern layer is the core of the knowledge graph. In the pattern layer, the refined knowledge is stored. Ontology model is built to express the knowledge of cybersecurity events and attack behaviours. Based on the ontology to establish pattern layer of the knowledge graph model for cybersecurity logs, the most basic thing is to establish the concept set.
The pattern layer of the knowledge graph model for cybersecurity logs is defined as formula (1).

$$\Omega = \{\mathbb{P}, \mathbb{E}, \Delta, \varphi, \emptyset, F, T, A, C, I\}$$  \hspace{1cm} (1)

In formula (1), $\mathbb{P}$ and $\mathbb{E}$ are two disjoint sets, the elements in $\mathbb{P}$ are defined as log records, and the elements in $\mathbb{E}$ are defined as cybersecurity events. $\Delta$ is a subset of $\mathbb{P}$, which is defined as high risk cybersecurity events. $\varphi$ is a function, which is used to classify cybersecurity events into different levels. Rule set $\emptyset$ is defined as recognition rules and analytical rules, which contains the rules required for ontology, using the appropriate logical language. Set $F$ is defined as entity classification, which contains security management centre, security computing environment, security area boundary and security communication network.

Set $T$ is defined as system type, which contains classified level ($T_1$), technology type ($T_2$) and service type ($T_3$). So that system type is defined as the set of $T = \{T_1, T_2, T_3\}$.

Set $A$ is defined as system applicability, which contains usage frequency ($A_1$), recovery time requirements ($A_2$) and unavailable loss size ($A_3$). So that system applicability is defined as the set of $A = \{A_1, A_2, A_3\}$.

Set $C$ is defined as system confidentiality, which contains data type ($C_1$), data leakage loss ($C_2$) and social impact of data leakage ($C_3$). So that system confidentiality is defined as the set of $C = \{C_1, C_2, C_3\}$.

Set $I$ is defined as system integrity, which contains integrity check required ($I_1$), scope of impact of integrity damage ($I_2$) and loss of integrity damaged ($I_3$). So that system integrity is defined as the set of $I = \{I_1, I_2, I_3\}$.

Through classified protection data contains a lot of accurate information of information system, such as device type, operating system version, IP address, and so on. So that data gain can be acquired by these classified protection data and the cybersecurity logs, and more abundant and multi-dimensional logs can be obtained, which lays a good foundation for subsequent logs analysis and logs audit. And pattern layer diagram of the knowledge graph model for cybersecurity logs is shown in Figure 2.

**Figure 2.** Pattern layer diagram of the knowledge graph model for cybersecurity logs

In Figure 2, log recognition is mainly aimed at the logs of syslog. The types and formats of syslog logs are very complex. In order to ensure the efficiency of the cybersecurity logs, it is necessary to identify and classify the types of logs in advance, and then add data benefits by combining recognition rules and analytical rules.

According to the types of logs, the classified protection information gain is to associate the corresponding classified protection data, gain the data of the cybersecurity logs, supplement the relevant information, such as the corresponding device type, open port, operating system version, device IP address, location, and so on. The information is supplied, such as the classified protection level of the information system, the technology type of the information system, location, and so on. At the same time, GeoIP and other tools can be used to obtain the country where script is located, ASN and so on.
Compliance analysis can be used by detection rules to determine the alarm of log, such as login attempt fails. And compliance rules can be used to determine the compliance items reflected in the logs. So that the accumulated knowledge of cybersecurity logs can be continuously transformed into algorithm model so as to form a knowledge graph.

4. Structure of the Knowledge Graph for Cybersecurity Logs

The construct of the ontology and knowledge can be stored in the neo4j graph database. Servers, network devices and security devices can be analysed as nodes. The nodes can have many attributes, and the classified protection are important attributes of the nodes. And the nodes can be grouped according to the security management centre, security computing environment, security area boundary and security communication network, so that the entities can be extracted from cybersecurity logs.

In order to audit and analyse security events from massive logs data and trace the origin of events from the cybersecurity logs, the knowledge graph for cybersecurity logs is built by knowledge extraction and knowledge fusion. Knowledge extraction is formed by entity extraction, relation extraction and attribute extraction. Knowledge fusion is formed by entity alignment, ontology construction and data gain. Classified protection data is integrated in entity alignment, and classified protection information gain is used in ontology construction. The structure of the knowledge graph for cybersecurity logs is shown as Figure 3.

**Figure 3.** The structure of the knowledge graph for cybersecurity logs

In Figure 3, the cybersecurity logs can be processed whether it is different sources or formats, so that the knowledge extraction of the cybersecurity logs can be realized. In addition, the classified protection data can be integrated to support entity alignment of real-time and historical logs of all devices, and association analysis can be realized across multiple data sources.

Through knowledge extraction and knowledge fusion, ontology construction and data gain are carried out for the structured fields, so that the knowledge graph for cybersecurity logs is built. The logs from the scattered event sources can be associated, and cybersecurity logs can be comprehensively and uniformly analysed and processed. By using classified protection information gain, the attack scenario can be reconstructed so as to reduce the false alarm rate.

By building the knowledge graph for cybersecurity logs, variety logs of cybersecurity equipment and security settings can be supported to analyse, and the personnel workload can be reduced, the efficiency of the system to find abnormal events can be greatly improved. The cybersecurity logs can be audited comprehensively. The model can support strong big data processing and can support PB level data.

5. Conclusion

For the complexity and massive of cybersecurity logs, it is difficult to build an accurate model based on prior knowledge in the face of big data logs, which means that big data analysis needs huge computing time and space cost, and the exact solution may not be able to be obtained within the time of acceptance.
The knowledge graph model for cybersecurity logs based on ontology and classified protection can be comprehensively audit the logs of various network devices, security devices, security systems, host operating systems, databases and various application of important information systems, and the logs of PB level can be supported to analysis by using big data technology.

The knowledge graph model for cybersecurity logs is built based on ontology and classified protection, the model has heuristic characteristics and generalization abstract ability, it is high self-organization and self-adaptability, so that it can be directly analysed and processed on the data without the need for precise modelling of the problem. It is suitable for big data analysis of large-scale cybersecurity logs. And an effective method of cybersecurity logs for audit and analysis is provided.

Model selection and statistical analysis method is used in traditional logs analysis, but this often relies on domain knowledge alone to perceive threats. The defects of domain knowledge are heavily dependent on expert knowledge, while expert knowledge is limited. The knowledge graph model for cybersecurity logs can audit and find unknown vulnerabilities, and provide multi-level reports. So that the security situation can be monitor from different perspectives, and respond can be deal with in time.

6. Acknowledgments
This research was financially supported by the National Key R&D Program of China (2018YFB0803503).

7. References
[1] Muhammad Mudasar Yamin, Basel Katt, Vasileios Gkioulos. Cyber ranges and security testbeds: Scenarios, functions, tools and architecture [J]. Computers & Security, 2020(88): 1-26.
[2] LIU Yu-Hua, ZHANG Ru-Min, ZHANG Jing-Yu, et al. Semantic-Enhanced Visual Abstraction of Large-Scale Multivariate Graphs [J]. Chinese Journal of Computers, 2020, 43(1):136-150.
[3] XIAO Ding, WANG Qian-Yu, CAI Ming, et al. Research on Implicit Interference Detection Based on Knowledge Graph in Smart Home Automation[J]. Chinese Journal of Computers, 2019, 42(6):1190-1204.
[4] REN Fei-Liang, SHEN Ji-Kun, SUN Bin-Bin, et al. A Review for Domain Ontology Construction from Text [J]. Chinese Journal of Computers, 2019, 42(3):654-676.
[5] TAO Yuan, GUO Fengming, LUO Zheng, et al. GA/T 1349-2017 Information Security Technology- Knowledge Library Interface Specification for Cybersecurity Classified Protection. Beijing: Standards Press of China, 2017.
[6] MA li, CHEN Gongyong, ZHANG Zhenfeng, et al. GB/T 22239-2019 Information Security Technology- Baseline for Classified Protection of Cybersecurity. Beijing: Standards Press of China, 2019.
[7] CHEN Gongyong, LI Ming, LI Shuilin, et al. GB/T 28448-2019 Information Security Technology- Evaluation Requirement for Classified Protection of Cybersecurity. Beijing: Standards Press of China, 2019.
[8] JIANG Yong, LI Chao, LI Quixiang, et al. GB/T 25070-2019 Information Security Technology-Technical Requirement of Security Design for Classified Protection of Cybersecurity. Beijing: Standards Press of China, 2019.
[9] HUO Shanshan, REN Weihong, Liu Jian, et al. GB/T 36958-2018 Information Security Technology- Technical requirements of security management center for classified protection of cybersecurity. Beijing: Standards Press of China, 2019.
[10] LIAO Xiangke, LI Shanshan, DONG Wei, et al. Survey on Log Research of Large Scale Software System [J]. Journal of Software, 2016, 27(8): 1934-1947.
[11] TAO Yuan, ZHANG Yuxiang, MA Siyuan, et al. Combining the Big Data Analysis and the Threat Intelligence Technologies for the Classified Protection Model [J]. Cluster Computing, 2017, 20 (2):1035-1046.
[12] CHEN Xingshu, ZENG Xuemei, WANG Wenxian, et al. Big Data Analytics for Network Security and Intelligence [J]. Advanced Engineering Sciences, 2017, 49 (3):1-12.
[13] WANG Wenjie, HU Baiqing, LIU Chi. A Survey of Open Source Softwares for Big Data Governance and Security [J]. Netinfo Security, 2017, 17(5):28-36.
[14] LIU Yadong, HU Dewen. High Performance Computing in the Viewpoint of Brain Research [J]. Chinese Journal of Computers, 2017, 40 (9): 2148-2166.

[15] Jürgen Schmidhuber. Deep Learning in neural networks: An Overview [J]. Neural Networks. 2015, 61: 85-117.

[16] PAN Quan, HU Yumei, LAN Hua, et al. Information Fusion Progress: Joint Optimization Based on Variational Bayesian Theory [J]. Acta Automatica Sinica, 2019, 45 (7): 1207-1223.