Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 7875910, 12 pages
https://doi.org/10.1155/2022/7875910

Research Article

An Automatic Assessment Method of Cyber Threat Intelligence Combined with ATT&CK Matrix

Shuqin Zhang, Peng Chen, Guangyao Bai, Shijie Wang, Minzhi Zhang, Shuhan Li, and Chunxia Zhao

College of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, China

Correspondence should be addressed to Shuqin Zhang; zhangsq@zut.edu.cn and Peng Chen; chenpeng@zut.edu.cn

Received 28 June 2022; Accepted 28 July 2022; Published 8 August 2022

Academic Editor: Zhiguo Qu

With the occurrence of cyber security incidents, the value of threat intelligence is coming to the fore. Timely extracting Indicator of Compromise (IOC) from cyber threat intelligence can quickly respond to threats. However, the sparse text in public threat intelligence scatters useful information, which makes it challenging to assess unstructured threat intelligence. In this paper, we proposed Cyber Threat Intelligence Automated Assessment Model (TIAM), a method to automatically assess highly sparse threat intelligence from multiple dimensions. TIAM implemented automatic classification of threat intelligence based on feature extraction, defined assessment criteria to quantify the value of threat intelligence, and combined ATT&CK to identify attack techniques related to IOC. Finally, we associated the identified IOCs, ATT&CK techniques, and intelligence quantification results. The experimental results shown that TIAM could better assess threat intelligence and help security managers to obtain valuable cyber threat intelligence.

1. Introduction

With the development of technologies, the scale of online devices has reached 6 billion [1]. Although the Internet has facilitated people's lives in various aspects, it cannot be ignored that the risks of information exposure during data transmission are increasing day by day [2]. Attackers may explore cyber vulnerabilities, launch attacks to obtain private information, and finally conduct malicious attacks [3]. Therefore, it is vital to protect the privacy of users in a network environment [4]. Several traditional defence methods against network have existed: firewalls, system patches, authentication, information encryption, and intrusion detection systems, etc. Since cyber attacks are becoming more sophisticated, traditional security protection strategies based on passive defence measures are difficult to deal with various types of attacks. The main reasons are as follows: (i) Vulnerabilities are unpredictable, and the attack methods used by attackers are constantly changing. It is difficult to use a general method to deal with network attacks. (ii) For advanced persistent threats, feature detection-based protection technology has been failed, and traditional means have been unable to cope. Therefore, it is imperative to strengthen the cyber security capabilities based on vulnerability analysis and cyber threat intelligence (CTI) information extraction and improve the network's active security defence performance [5]. CTI is able to describe the attack behavior, provide the context of the network attack, and guide how to defend against the attack, which can play a crucial role in network security protection.

IOC (Indicator of Compromise) describes the behavioral characteristics of cyber threats, including static information (such as signatures), and dynamic characteristics (such as the behaviors that malware takes on the victim's computer). Meantime, it reveals the attack strategy adopted by attackers, and the strategy can be used to match existing network threats and discover variants or similar cyber-attack cases [6]. Existing CTIs usually come in the form of IOCs. Once these IOCs are collected by threat intelligence platforms and formatted according to threat information sharing standards such as Trusted Automated Exchange of Intelligence Information (TAXII), they can be automatically converted

and imported into various defence mechanisms such as intrusion detection systems. The virtue of threat intelligence collection resides in that data used as a source of threat alerts can be extracted from cyber open source intelligence (OSINT) [7] based on the specific demand of the organization. Due to the uncertainty of threat intelligence data sources, the same category of threat intelligence might exist from multiple sources; hence, their quality and credibility must be assessed to avoid "data poisoning." Reliability and quality are the key assessment factors of OSINT.

ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) is a knowledge base that reflects the attack life cycle and attack behavior. ATT&CK organizes the adversary's tactics and techniques through real-world observations. It can analyze the TTPs potentially used to execute an attack from the attacker's perspective and displays them in the form of a matrix [8]. Threat intelligence provides insight into attackers and their malicious activity, including context, indicators, and operating recommendations. Both ATT&CK and threat intelligence involve the study of attacker information. Therefore, we apply the ATT&CK matrix to threat intelligence assessment.

Extracting threat information from CTI reports faces two main difficulties: (i) Since CTI reports are written in natural language, extracting structured attack behaviors requires analyzing the content in unstructured CTI text (i.e., sparse text). (ii) Attack information is scattered in the report, which makes it difficult to fully analyze the techniques attackers may use. For the above two reasons, we propose TIAM, a model for automatically assessing CTI. The contributions of this paper are mainly reflected in three aspects:

(i) Propose a quality assessment model of threat intelligence, which processes highly sparse text and automatically calculates the score of intelligence

(ii) Propose an automatic threat classification method based on feature extraction, which takes dictionary word density and feature word weight as classification criteria

(iii) Correlate the ATT&CK attack matrix with CTIs for associating threat intelligence with attacking organizations. Experiments show TIAM can automatically identify the tactics and techniques used by the attacker

In this report, we first discuss existing methods and related research work in Section 2, describe the currently widely used threat intelligence sharing standards and platforms in Section 3, and then propose a new threat intelligence assessment methodology in Section 4. Subsequently, the relevant experimental verification and results analysis are presented in Section 5. Finally, Section 6 provides the concluding remarks.

2. Related Work

CTI plays an important role in cyber security. However, most of the threat intelligence is inaccurate, incomplete, and out of date. Therefore, low-quality CTI is identified to be a pivotal issue [9]. In contrary, high-quality CTI can promote identifying potential threats at early stage or even prevent network attacks.

The quality of threat intelligence can be assessed from two aspects: threat intelligence quantification and threat information extraction, which will be discussed in detail in the following.

2.1. Quantification of Threat Intelligence. Large numbers of works have been done on threat intelligence quality assessment. Some scholars quantified threat intelligence from the user's perspective. Li et al. [10] introduced five-dimension criteria, namely, availability, reliability, usability, relevance, and presentation quality, and formulated a hierarchical data quality framework from the perspective of data users. They constructed a dynamic big data quality assessment method based on a feedback mechanism. Jaikrit et al. [11] proposed a quality assessment model for Internet products and services. The model divided the assessment indicators into two categories, which were those that met the design requirements and those that exceeded the requirements, and introduced several Internet-related assessment factors. Qiang et al. [12] proposed a threat intelligence assessment framework, which realized a multidimensional quantitative assessment of threat intelligence services from the perspective of users.

Other scholars attempted to assess the quality of threat intelligence from its source. Andrea et al. [13] implemented an automated scheme to quantify the CTI sources to understand the relevance between various sources of cyber threats and proposed a signature-driven approach to assess OSINT sources. Schaberreiter et al. [14] utilized quantitative parameters to assess cyber threat information sources. They introduced such parameters as extensiveness, verifiability, false positives, and intelligence and proposed a method for quantitative assessment.

Furthermore, some scholars assessed the quality of threat intelligence directly. Magee et al. [15] developed a threat intelligence collection system to identify threats by classifying threat intelligence according to its type, maliciousness, and credibility. Botega et al. [16] proposed IQESA for assessing information quality. This strategy consisted of three stages: capturing data and information quality requirements, defining metric functions and quantifying quality dimensions, and instantiating contextual information using ontology.

However, while assessing the quality of intelligence, the above works only considered the information contained in the intelligence, but ignored the potential attack techniques and tactics that attackers may use in threat intelligence.

2.2. Threat Information Extraction. Recently, attack behavior extraction from CTIs has attracted attention from industry and academia and is considered an effective method to defend against network attacks [17]. NLP (natural language processing) technology, an increasingly mature tool, is used by some scholars to analyze CTI from multiple aspects. Liao et al. [18] presented iACE to collect threat intelligence and

implement the automatic extraction of IOC. iACE can automatically locate the IOC token and context and analyze their relationship. Husari et al. [19] developed a TTPDrill tool, which combined NLP and information retrieval to automatically mine threat behaviors, and constructed STIX-formatted TTPs from threat reports. Zhou et al. [20] designed CTI View, a system that used NLP technology to automatically process CTI. CTI View can automatically extract and analyze CTI text information released by security vendors.

Other scholars applied machine learning technology to information extraction. Mulwad et al. [21] used the support vector machine (SVM) classifier to identify potential vulnerability description and then used the taxonomy in Wikipedia to extract vulnerabilities and attacks. But this classifier only identified and extracted two kinds of concepts, one is the attack means, and the other was the consequences. Yuma et al. [22] proposed a method to automatically generate interpretable IOCs by tracking malware processes. The main idea was to enumerate the key information of all potential IOCs, then continuously optimized and combined this information to maximize the interpretability and accuracy of threat intelligence, and finally generated reliable IOCs. Lv et al. [23] proposed a threat intelligence analysis method based on the detection of the attack chain and network traffic. The method detected the network anomaly traffic, analyzed the relationship of characters, and combined the characteristics of each stage of the kill-chain to restore the process of attack.

Besides, some scholars applied deep learning-related technologies such as the long-short-term memory (LSTM) [24] network, the convolutional neural network (CNN) [25], and the recurrent neural networks (RNN) [26] to information extraction. Luo et al. [27] proposed EDL-WADS, a deep learning-based network attack detection system that used all three models of CNN, MRN, and LSTM for network attack detection. Yan et al. [28] used two techniques of feature weighting and BERT-BiGRU and proposed a method to analyze IIoT threat intelligence, which could realize automatic identification of attack behavior and attack strategies. Wang et al. [29] improved the PCNN-ATT model and proposed a DRL-ETPCNN-ATT-based method for remote supervised relation extraction, which could extract threat intelligence from unstructured text.

However, these works mentioned above were limited to the extraction of threat intelligence and ignored the quality assessment of extracted information.

3. Threat Intelligence Sharing Standards and Platforms

IoT devices can generate a large amount of sensitive and private information while working, which is extremely attractive to attackers [30]. The sharing of CTI is an effective measure to strengthen cyberspace security collaboration and improve cyber security. The threat intelligence community can leverage this information to better understand the situation and share intelligence with communities, organizations, and the public.

3.1. Threat Intelligence Sharing Standards. Today's network attack behaviors are becoming increasingly complex and attack patterns are constantly evolving, which lead to the requirement of automatic information processing, and rapid sharing and responding to changes in cyber attacks [14]. Since threat indicators can specify information such as threat actors, vulnerabilities exploited, attack programs, and threat activities related, a unified threat indicator becomes a prerequisite for automation. STIX is a standardized language widely used to represent cyber threat information.

STIX is used as a serialization format to exchange CTI, and it is one of the most widely used threat intelligence sharing languages [31]. STIX can describe various characteristics of threat intelligence, such as threat signatures, threat activities, and security incidents. It increases the threat intelligence exchanging efficiency and accuracy, improves the responsiveness of security managers to threats, and helps organizations effectively realize the automation of cyber threat management and application. Given that, we convert the obtained threat intelligence into STIX format. The data conversion method will be introduced in Section 4.1 as an example.

3.2. Threat Intelligence Sharing Platforms. Dandurand [32] proposed that a CTI sharing platform should include three aspects: (i) enabling information sharing, (ii) automating information exchange, and (iii) facilitating the generation and updating of threat intelligence data. Today, a growing number of threat intelligence sharing platforms have begun to enter the public's vision. In Table 1, some threat intelligence sharing platforms are listed.

Some threat intelligence sharing platforms can not only realize the sharing of information but also conduct online analysis and detection of threat intelligence text, identify the IOC information contained in the text, and then give relevant warning information. For example: Qi-Anxin and ThreatBook, we compare them with our model in Section 5.1 of the paper.

4. Cyber Threat Intelligence-Automated Assessment Model (TIAM)

How to assess the quality of a large scale of threat intelligence from many enterprises and organizations has become a key research problem. The vector space is a simple and effective text representation model, generally used in discrete text [33]. The vector representation is characterized by high dimensionality and sparsity due to the nature of text. The vector space contains feature words extracted from a large amount of text. If the feature words in the vector space do not exist in the corpus, the value of feature word is set to 0; otherwise, the value is set to the number of times it appears. Figure 1 presents one example of word vector representation for sparse text.

The model proposed in this paper is called Cyber Threat Intelligence-Automated Assessment Model (TIAM), which can realize automatic assessment of CTI. TIAM analyzes the content in CTI by extracting IOC with a specific format,

Table 1: Some threat intelligence sharing platforms.

Threat intelligence platforms	Description	Links
Malware Information Sharing Platform (MISP)	It collects and stores network security indicators and threat information and can analyze malware and network security events.	https://github.com/MISP/MISP
Qi-Anxin Threat Intelligence Center	It has the capability to analyze threat intelligence such as discovering major threats, providing context for decision-making in response to incidents, and providing security early warning.	https://www.qianxin.com/threat/reportaptlist
ThreatBook Intelligence Community	It includes functions such as APT tracking, sample Trojan analysis, and vulnerability analysis.	https://x.threatbook.cn/
Facebook Threat Exchange	The Threat Exchange platform contains security information on malicious links, phishing websites, unwanted software, and network attacks.	https://github.com/facebook/ThreatExchange
IBM X-Force Exchange	IBM X-Force Exchange is a threat intelligence cloud platform that enables rapid sharing of threat intelligence.	https://exchange.xforce.ibmcloud.com/
NSFOCUS	It conducts research on the world's latest security vulnerabilities, unearths hidden information, and publishes security research reports.	http://www.nsfocus.net/index.php?act=sec_bug
Alien Vault Open Threat Exchange (OTX)	This is a public-facing threat intelligence sharing community where participants can obtain the latest threat information and update their defence systems by downloading the latest threat information through an API interface.	https://www.alienvault.com
Eclectic IQ	It is an extensible and open platform that combines many front-line skills to automate threat intelligence processing.	https://www.eclecticiq.com

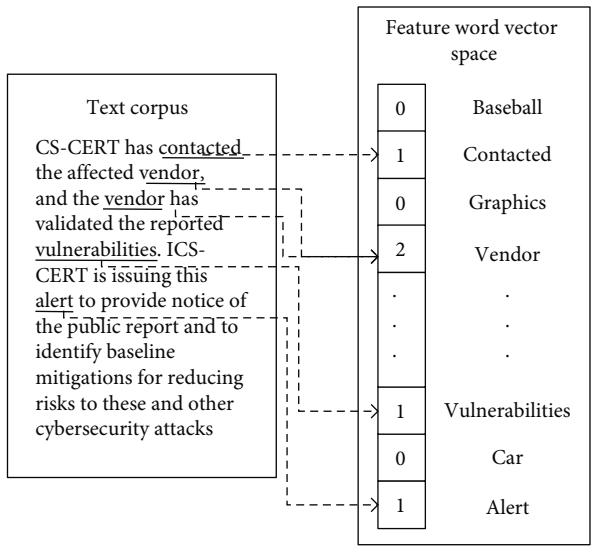

Figure 1: Word vector representation for sparse text.

which can help security managers assess the value of CTI. IOC can assist security managers in finding specific types of characteristic data in system or network logs, which leads to the identification of infected targets. These characteristic data include IP address, domain name, malicious file hash associated with C2 servers (Command and Control Server), malware download, and other characteristics. In the process of CTI assessment, TIAM is combined with the ATT&CK matrix, which can help security managers analyze attack techniques an attacker might use and then find mitigations to prevent it. Figure 2 presents the TIAM framework, which consists of TICA, TIE, and TIQA modules.

TICA is responsible for collecting and aggregating threat intelligence, converting them to STIX format, and storing them into a database. TIE classifies sparse text-based threat intelligence, identifies, and filters IOCs. TIQA correlates the IOCs extracted from TIE with the intelligence collected by TICA and also assesses the intelligence quality and identifies the ATT&CK attack techniques related to the assessed intelligence.

4.1. TICA Module. TICA is primarily composed of two parts: data collection and database. Data collection is an automated data collection system, which collects and converts community-sourced threat intelligence (e.g., CVE [34], CWE [35], and CAPEC [36]), security website source intelligence (e.g., Lvmeng and X-Force), and open-source intelligence (e.g., ATT&CK knowledge base) to STIX format.

TIAM can dynamically assign intelligence collection tasks to a python agent that crawls data based on the corresponding data collection environment. TIAM converts the data into STIX format and stores them in the database. We show a partial mapping of ATT&CK concepts to STIX objects in Table 2. Through data normalization, different databases can be connected to each other.

4.2. TIE Module. TIE categorizes intelligence obtained from various sources into IOC and non-IOC intelligence. The information obtained from various sources is of uneven quality, some of which may not be relevant to CTI. For example, some data only contains information related to product advertisements and news, which can be defined as non-IOC intelligence.

4.2.1. Threat Intelligence Categorization. For distinguishing IOCs from non-IOCs, TIE considers the feature word weight

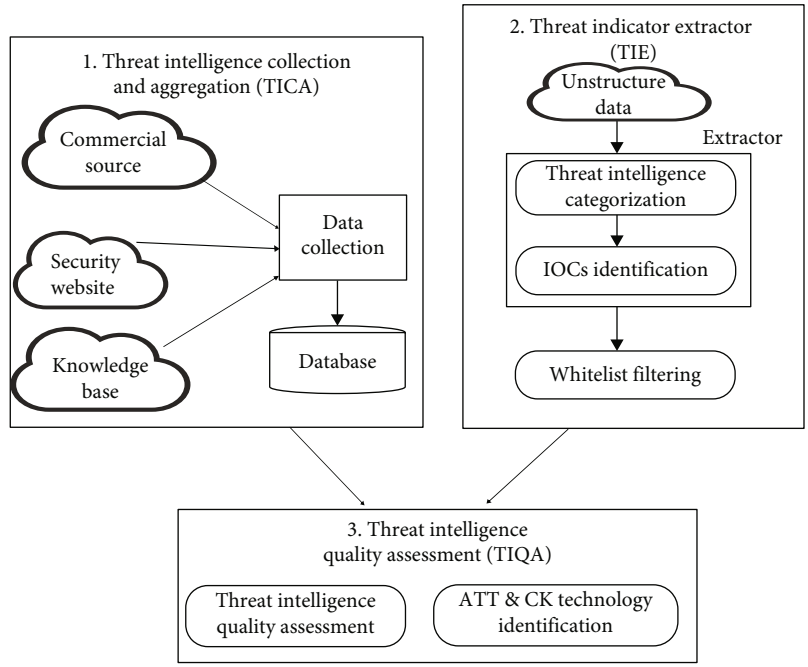

Figure 2: TIAM Framework. It consists of TICA, TIE, and TIQA modules.

Table 2: A mapping of ATT&CK concepts to STIX 2.0 objects.

ATT&CK concepts	STIX object	Custom type?
Tactic	x-mitre-tactic	Yes
Matrix	x-mitre-matrix	Yes
Mitigations	Course-of-action	No
Groups	Intrusion-set	No
Malicious	Malware	No
Software	Tool	No
Technique	Attack-pattern	No
Subtechnique	Attack-pattern where x_mitre_is_subtechnique = true	No
Procedure	Relationship where relationship_type = "uses" and target_ref is an attack-pattern	No

and non-dictionary word density as classification criteria as follows.

(1) Feature Word Weight. To ensure quick and efficient classification performance. The text needs to be transformed into an intermediate form, thus filtering out redundant and irrelevant features. In text vectors, feature values are usually used to represent the weights of feature words, which also reflect their importance. TIE calculates the weight of feature words through the TextRank [37], as shown in

$$WS(V_i) = (1-d) + d * \sum_{V_j \in \mathrm{In}(V_i)} \frac{w_{ji}}{\sum_{V_k \in \mathrm{Out}(V_j)} w_{jk}} WS(V_j), \tag{1}$$

where w_{ji} represents the weight from node i to node j. For a given vertex V_i, $\mathrm{In}(V_i)$ represents the set of vertices that point to V_i, $\mathrm{Out}(V_j)$ represents the set of vertices out of V_j, and d is a damping factor that can be set between 0 and 1.

(2) Non-dictionary Word Density. Dictionary words are common English words in dictionaries. IOC text contains information related to cyber threat behavior characteristics and the analysis codes of malicious samples; thus, it involves plenty of non-dictionary words. TIE uses the method proposed in [38] to calculate the non-dictionary word density.

4.2.2. IOC Identification. As the information contained in threat intelligence is unstructured, traditional natural language processing techniques have difficulty in identifying IOCs. Our study reveals that most of the information in threat intelligence shows a certain structure, such as malicious IP (xxx.xxx.xxx.xxx) and CVE number (CVE-xxxx-xxxx), which can be matched by regular expressions. The ability to match substrings in preprocessed text using regular

Table 3: IOCs' regular expressions and example.

IOC feature	Regex expression	Example
IP	^((25[0-5]|2[0-4]\d|[01]?\d\d?)\.){3}(25[0-5]|2[0-4]\d|[01]?\d\d?)$	41.208.110.46
Hash	\b[a-fA-F\d]{32}\b|\b[a-fA-F\d] {40}\b|\b[a-fA-F\d]{64}\b	830a09ff05eac9a5f42897ba5176a36a
Domain	\b(([a-zA-Z]\?*]|[\?*a-zA-Z][\?*a-zA-Z0-9\-]*[\?*a-zA-Z0-9)\.)+([A-Za-z0-9][A-Za-z0-9\-]*[A-Za-z0-9]|[A-Za-z0-9])\b	http://gezelimmi.com
URL	\b([a-z]{3,}\:\/\/[\S]{16,})\b	http://www.secureworks.com/
Email	\b([a-z][_a-z0-9-.]+@[a-z0-9-]+\.[a-z]+)\b	service@santander-sm.co.uk

Table 4: Assessment features and corresponding scores.

Assessment feature	Description	Attributes	Score
Alert	The possible harm through the IOCs	Marked high-risk in the database	3
		Marked medium-risk in the database	2
		Marked low-risk in the database	1
Created time	Timestamp related to IOC	Last day	5
		Last week	4
		Last month	3
		Last year	2
		Other	1
External reference	Other threat activities related to this indicator	Multi known reference	4
		Single known reference	3
		Unknown reference	2
		No reference	1
CVE	Check if the CVE is found in the extracted IOCs, and if so, check the CVSS	CVE with critical CVSS	5
		CVE with high CVSS	4
		CVE with medium CVSS	3
		CVE with low CVSS	2
		No CVE or CVE with no CVSS	1

expressions is available in many applications [39]. Therefore, TIE constructs some regular expressions [40] for different types of IOCs (Table 3).

TIE uses a whitelist to filter the obtained IOCs. Diverse typical whitelists are applied in the security industry, such as Alexa top 100W, Google Safe Browsing, and Cisco Umbrella. TIE uses the whitelist filter to process the received IOCs. If the IOC is found in the whitelist, then the intelligence information record for that IOC will be filtered out.

4.3. TIQA Module. TIQA takes IOCs as an input for assessing threat intelligence and correlates with TICA to analyze possible attack techniques.

4.3.1. Threat Score. By viewing the threat description, it is found that cyber-attacks that can cause great damage typically utilize multiple vulnerabilities, most of which are marked as high-risk vulnerabilities by CVE, while those exploit single and less threatening ones will cause less harm [41]. The threat scoring follows three principles: (i) the closer the attack time is to the threat intelligence release time, the less likely the attack will fail and the more damage it will cause; (ii) the more threat alerts are generated in an attack, the greater damage is caused to the system; and (iii) the more external threat activities are associated with a threat entity, the more significant impact will be exerted on the enterprise. Therefore, the threat can be quantified by exploring the attack time, the number of alerts, and the number of threat activities related to the threat entity. Threat intelligence score is a good measure of the quality of information in threat intelligence.

TIQA quantifies threat intelligence based on the above three threat features. It assigns values to the obtained IOCs (Table 4) and assesses the quality of threat intelligence with a scoring function defined as

$$TIS = \sum_{i=1}^{m} \sum_{j=1}^{3} \frac{s_{ij} \times w_{ij}}{m}. \quad (2)$$

Here, m represents the number of IOCs extracted from one threat intelligence, s_{ij} represents the score of j-th threat feature in i-th IOC, and w_{ij} is the weight of s_{ij}. The threat

Table 5: Some threat attack-related indicators.

Indicator_type	Indicator_value
MD5	dc63b4b9ee2f8486b96ce62be4a31e041d422ef7
Host	www.researchbundle.com
MD5	a75fdd9e52643dc7a1790c79cbfffe9348f80a9b0984eafd90723bf7ca68f4ce
CVE	CVE-2010-3333
MD5	b80d436afcf2f0493f2317ff1a38c9ba329f24b1
MD5	ed6ad64dad85fe11f3cc786c8de1f5b239115b94e30420860f02e820ffc53924
MD5	25ac3098261df8aa09449a9a4c445c91321352af
MD5	e547e8a8bc27d65dca92bc861be82e1c94b9c9aca8a2b75381e9b16e4ad89600
CVE	CVE-2012-0158
Host	www.viprambler.com

score (TIS) ranges from 1 to 10 ($1 \leq \mathrm{TIS} \leq 10$), and the higher the TIS value, the higher the quality of threat intelligence.

When a CVE vulnerability is published, it will include information about the vulnerability description, date, and comments. CVSS (Common Vulnerability Scoring System) [42] is an open standard for measuring the impact of vulnerabilities and can assess the severity of vulnerabilities. TIQA places the extracted complete CVE number into CVSS to assess the threat level and subsequently assigns score according to Table 4. The threat level caused by exploiting the vulnerability can be calculated with

$$\mathrm{Impact} = \lambda \mathrm{VL}[1 - (1 - \mathrm{Con}_a)(1 - \mathrm{Int}_a)(\mathrm{Ava}_a)]. \quad (3)$$

Among them, λ is the correction factor with a value of 10.41; VL is the difficulty of utilizing vulnerability, which is divided into four grades: critical, high, medium, and low. $C = (\mathrm{Con}_a, \mathrm{Int}_a, \mathrm{Ava}_a)$ indicates the confidentiality, integrity, and availability hazards caused by the vulnerability to the system.

TIQA calculates the weight of s_{ij} with

$$w_{ij} = \frac{t_{ij} + v_{ij} + \sum_0^n p_n}{\max\left(t_{ij} + v_{ij} + \sum_0^n P_n\right)}, \quad (4)$$

where t_{ij} denotes the score of IOC creation time, v_{ij} denotes the number of threat activities scores, $\sum_0^n p_n$ denotes the score of alerts, and n is the number of alerts.

4.3.2. Related ATT&CK Attack Technology. ATT&CK abstractly describes a framework composed of sequential attack tactics, each of which covers abundant attack techniques [43]. ATT&CK framework can help organizations predict the adversary's attack behavior, gain a comprehensive understanding of the attack techniques that attackers may use, and provide mitigation measures.

TIAM uses TIE to parse threat intelligence and extract IOCs. Then, TIQA correlates the extracted IOCs with ATT&CK knowledge base in TICA and automatically identifies attack techniques related to IOCs. Identifying attacks in threat intelligence can help security managers gain a more complete picture of the attack event.

5. Experiments and Assessment

5.1. Assessing One IOC Intelligence as an Example. We analyzed an APT attack on India's cyber space reported by Indian Infosec consortium [44]. First, 50 articles from an existing dataset were manually selected to label entities, and the occurrences of each entity were counted. Then, the top 800 entities with the most occurrences were selected as dictionary words. Finally, the feature words of the remaining articles were calculated and compared with the dictionary words. If a number of feature words appeared in the dictionary words, TIE considered the article to be threat intelligence. For those articles that were not considered threat intelligence, TIE used regular expressions to extract the IOCs and calculated the non-dictionary word density. If a large number of IOCs could be extracted and the density of non-dictionary words was high, TIE considered the intelligence to be threat intelligence as well.

TIAM forwarded this report into TIE, which removed useless stop words and characters. After feature word calculation, this report was considered threat intelligence. Performing the whitelist filtering, 15 out of 32 IOC indicators features, which extracted by TIE, related to threat attacks were finally identified, and 10 of these are presented in Table 5.

TIQA associated the first threat indicator in Table 5 with the database in TICA. After correlation analysis, the following were found: (i) this threat indicator was created in September 2021; (ii) five alerts were detected, namely, "recon_beacon," "persistence_auto," "allocates_rwx," "network_http," and "antivm_memory_available," and the database in TICA defines them as high-risk, high-risk, medium-risk, medium-risk, and low-risk, respectively; and (iii) this threat indicator was associated with three kinds of threat activities: attacks at US polling places, attacks against the US government, and yarex-related malware campaigns.

TIQA assigned values to the identified information according to Table 4 and repeated the above analysis process for the remaining threat indicators. Through the threat scoring function, this report had a final score value of 7.43.

In the process of association analysis, three alerts were reported through the database in TICA which were "persistence_autorun," "antivm_memory_available," and "recon_

Figure 3: One threat intelligence scoring and visualization. Green circle represents threat reports, pink circles represent threat indicators, blue circles represent ATT&CK techniques, and yellow circle represents threat report quality scores.

Table 6: Comparison of threat intelligence analysis platforms.

Platform	Number of alerts	Number of ATT&CK techniques	IOC created time	IOC-related external activities	Visual analysis
TIAM	18	9	√	√	√
ThreatBook	8	4	√	×	√
Qi-AnXin	8	0	√	×	√

beacon" and then were associated with the ATT&CK. TIQA found three attack techniques, namely, Registry Run Keys/Start Folder, Software Packing, and Automated Collection, which might be used by attackers for launching attacks.

Registry run keys/start folder:attackers add compromise code to the startup folder or use registry run keys to ensure persistence. The "run key" added to an entry when a user logs in will execute automatically [45]. It is possible for attackers to gain account-level privileges and execute the malicious code in the context of the user.

System information discovery:attackers can obtain hardware and operating system details such as patches, service packs, hotfixes, and architectures. Attackers can exploit this information to enhance their own operability, for example, to determine whether the target is fully infected [46].

Automated collection:after a successful implementation of the attack behavior, the attackers can automatically collect data inside the infected host. Attackers may use a script interpreter or command-line operations to search for information that matches the set criteria, including location, file type, and name [47].

Through analysis, it is found that these three techniques are in line with the characteristics commonly used in cyberattack against cyberspace. Figure 3 visualizes the threat

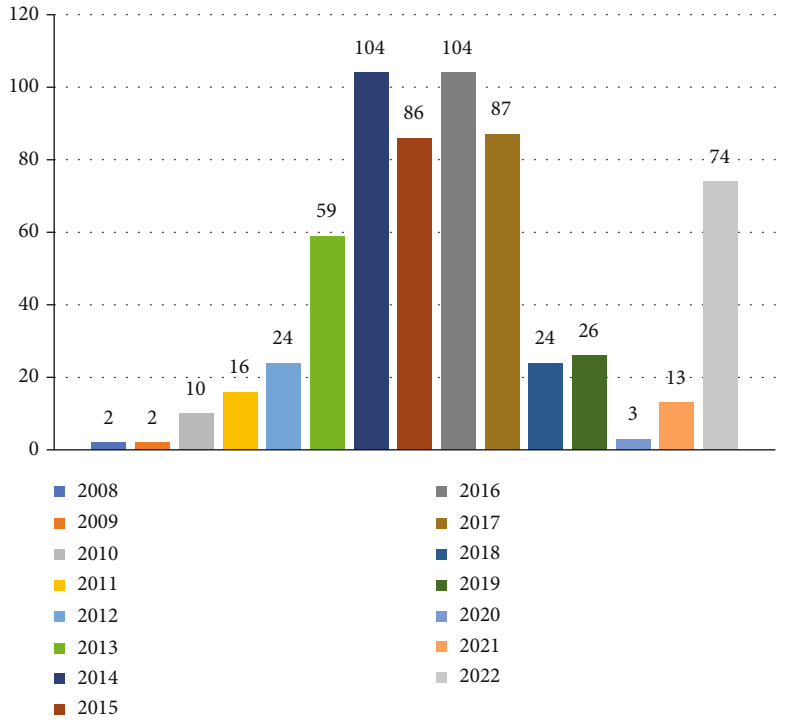

Figure 4: Annual threat intelligence statistics.

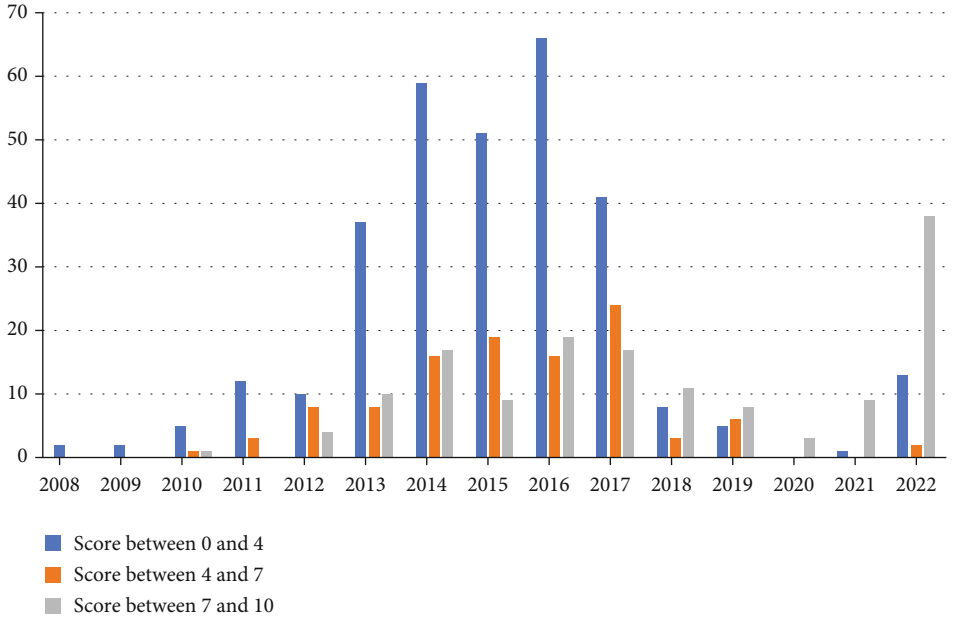

Figure 5: Quantitative assessment results of CTI for each year.

intelligence quality score, identified threat indicators, and the ATT&CK attack techniques.

We compared some of the threat intelligence platforms (ThreatBook, Qi-AnXin) in Table 1 against our model in five aspects: (i) the number of threat alerts generated, (ii) the number of ATT&CK attack techniques, (iii) whether focus on the time of IOC creation, (iv) IOC-related threat activity can be identified, and (v) whether to include visual analysis of results. The comparison results are shown in Table 6.

From Table 6, we can find that the TIAM proposed in this paper is superior compared with the other two in terms of overall performance. The reason for its advantage, we believe, is that TIAM integrates more vulnerability libraries and connects with ATT&CK. TIAM focuses on assessing

Table 7: Top 10 most frequently used ATT&CK techniques and occurrences.

Attack technique id and name	Occurrences in the reports
T1129 Shared Modules	872
T1027 Software Packing	831
T1082 System Information Discovery	527
T1053 Scheduled Task	469
T1057 Process Discovery	391
T1040 Network Sniffing	322
T1547 Registry Run Keys/Startup Folder	255
T1012 Query Registry	240
T1143 Hidden Window	192
T1055 Process Injection	184

the possible impact of each IOC in threat intelligence and quantifying threat intelligence, thus generating more comprehensive alert information, while the other two are more focused on analyzing the attacker's attack path, and network attack behavior. The reason why Qi-Anxin did not identify the ATT&CK attack techniques is that its database for file detection is not connected to ATT&CK.

5.2. IOC Intelligence Quality Assessment and Analysis. TIAM assesses threat intelligence from APTnotes [48], which contains APT reports over the past 13 years. Figure 4 shows the statistics on the volume of threat intelligence by year. Most of the reports are from the top security vendors, such as McAfee, Aurora, Kaspersky, SecureWorks, Cisco Talos, and FireEye. The reports can be grouped into three categories: analysis of APT organizations, analysis of the current raging malware, and analysis of a specific attack.

We used TIAM to assess APT reports and found that an average of 65 threat indicators appeared in each article and 15% of articles had more than 100 threat indicators. TIAM extracted 39,090 pieces of threat indicators from 632 reports. By quantifying the quality of these reports, TIQA found that there were 312 threat intelligence with a score between 1 and 4, 106 threat intelligence with a score between 4 and 7, 146 threat intelligence with a score between 7 and 10, and 68 threat intelligence identified as non-IOCs. Among them, 2016 contained the most low-scoring threat intelligence, while 2022 contained the most high-scoring threat intelligence. Figure 5 shows the quantitative assessment results of CTI for each year. Meanwhile, TIAM identified 5505 attack techniques (including recurring techniques) from these reports. Table 7 shows the 10 most frequently used attack techniques and the number of occurrences in the reports.

Shared Modules can instruct the window template loader using NTDLL.dll to load DLLs from arbitrary local and Universal Naming Convention (UNC) network paths. NTDLL.dll is included in the Windows Native API, which is called by functions such as CreateProcess and LoadLibrary of the Win32 API [49]. After analysis, we found that this technique was difficult to monitor with the current tools. Redundant DLL made monitoring and detection efforts pointless. Obviously, the commonly used techniques come from different tactics. In fact, this experimental result reflects our intuitive understanding of the attack lifecycle [50], where typical attacks must consist of a unique set of strategies that can be implemented with different techniques.

6. Conclusions

In order to tackle the problem of high sparsity and uneven quality of the information in threat intelligence, this paper proposes TIAM to analyze unstructured CTI, classify and identify threat intelligence through text features, and extract IOC information automatically. TIAM introduces the existing attack technology knowledge in ATT&CK into automated assessments of threat intelligence. In the end, it provides the quantitative assessment results of unstructured threat intelligence. In the future, we will expand the scope of information extraction, beyond IP, hash, URL, and features that have a fixed format and introduce technologies such as machine learning.

Data Availability

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

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the Key Program Research Fund of Higher Education of Henan, China (grant number No. 21A520053 and No. 19A520048).

References

[1] G. Rathee, S. Garg, G. Kaddoum, and B. J. Choi, "A decision-making model for securing IoT devices in smart industries," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4270–4278, 2021.

[2] L. Sun and J. Wu, "A scalable and transferable federated learning system for classifying healthcare sensor data," *IEEE Journal of Biomedical and Health Informatics*, vol. 1, pp. 1–1, 2022.

[3] S. G. Abbas, I. Vaccari, F. Hussain et al., "Identifying and mitigating phishing attack threats in IoT use cases using a threat modelling approach," *Sensors*, vol. 21, no. 14, p. 4816, 2021.

[4] L. Sun, R. Zhou, D. Peng, A. Bouguettaya, and Y. Zhang, "Automatically building service-based systems with function relaxation," *IEEE Transactions on Cybernetics*, vol. 52, pp. 1–14, 2022.

[5] G. W. Shen, W. L. Wang, Q. L. Mu, Y. Pu, Y. Qin, and M. Yu, "Data-driven cybersecurity knowledge graph construction for industrial control system security," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8883696, 13 pages, 2020.

[6] M. Conti, T. Dargahi, and D. Ali, "Cyber threat intelligence: challenges and opportunities," *Cyber Threat Intelligence*, vol. 70, pp. 1–6, 2018.

[7] G. Michagffel and J. K. Min, "Intelligence in the internet age: the emergence and evolution of Open Source Intelligence (OSINT)," *Computers in Human Behavior*, vol. 28, no. 2, pp. 673–682, 2012.

[8] K. Kim, Y. Shin, J. Lee, and K. Lee, "Automatically attributing mobile threat actors by vectorized ATT&CK matrix and paired indicator," *Sensors*, vol. 21, no. 19, p. 6522, 2021.

[9] F. B. Kokulu, A. Soneji, T. Bao et al., "Matched and mismatched SOCs: a qualitative study on security operations center issues," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1955–1970, London, United Kingdom, 2019.

[10] C. Li and Y. Zhu, "Big research data and data science," *Data Science Journal*, vol. 14, pp. 1–10, 2015.

[11] K. Jaikrit, C. J. Erick, F. N. Fiona, and R. B. Ram, "Information quality on the world wide web: development of a framework," *International Journal of Information Quality*, vol. 2, no. 4, pp. 324–343, 2011.

[12] L. Qiang, J. Zhengwei, Y. Zeming, L. Baoxu, W. Xin, and Z. Yunan, "A quality evaluation method of cyber threat intelligence in user perspective," in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pp. 269–276, New York, NY, USA, 2018.

[13] T. Andrea, R. Samuel, and M. Max, "A feature-driven method for automating the assessment of OSINT cyber threat sources," *Computers & Security*, vol. 113, p. 102576, 2022.

[14] T. Schaberreiter, V. Kupfersberger, K. Rantos et al., "A quantitative evaluation of trust in the quality of cyber threat intelligence sources," in *Proceedings of the 14th International Conference on Availability, Reliability and Security*, pp. 1–10, Canterbury, United Kingdom, 2019.

[15] J. C. Magee, A. M. Andrews, M. W. Nicholson et al., *Collective Threat Intelligence Gathering System*, vol. 1, p. 228, 2014.

[16] L. C. Botega, J. O. de Souza, F. R. Jorge et al., "Methodology for data and information quality assessment in the context of emergency situational awareness," *Universal Access in the Information Society*, vol. 16, no. 4, pp. 889–902, 2017.

[17] V. Mavroeidis and S. Bromander, "Cyber threat intelligence model: an evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence," in *2017 European Intelligence and Security Informatics Conference (EISIC)*, pp. 91–98, Athens, Greece, 2017.

[18] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah, "Acing the ioc game: toward automatic discovery and analysis of open-source cyber threat intelligence," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 755–766, Vienna, Austria, 2016.

[19] G. Husari, E. Al-Shaer, M. Ahmed, B. Chu, and X. Niu, "Ttpdrill: automatic and accurate extraction of threat actions from unstructured text of cti sources," in *Proceedings of the 33rd Annual Computer Security Applications Conference*, pp. 103–115, Orlando, USA, 2017.

[20] Y. H. Zhou, Y. Tang, M. Yi, C. Y. Xi, and H. Lu, "CTI view: APT threat int-elligence analysis system," *Security and Communication Networks*, vol. 2022, Article ID 9875199, 15 pages, 2022.

[21] V. Mulwad, W. Li, A. Joshi, T. Finin, and K. Viswanathan, "Extracting information about security vulnerabilities from web text," in *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pp. 257–260, Lyon, France, 2011.

[22] K. Yuma, O. Yuto, K. Yuhei, I. Makoto, and S. Koushik, "EIGER: automated IOC generation for accurate and interpretable endpoint malware detection," in *Proceeding of the 35th Annual Computer Security Applications Conference*, pp. 687–701, San Juan, USA, 2019.

[23] Z. P. Lv, Y. B. Zhong, and Z. J. Gu, "Threat intelligence analysis research based on kill chain and network traffic detection," *Application Research of Computers*, vol. 34, pp. 1794–1797, 2016.

[24] Y. L. Wang, L. Sun, and S. Subramani, "CAB: classifying arrhythmias based on imbalanced sensor data," *KSII Transactions on Internet & Information Systems*, vol. 15, no. 7, pp. 2304–2320, 2021.

[25] J. Huang, B. Chen, B. Yao, and W. He, "ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network," *IEEE Access*, vol. 7, pp. 92871–92880, 2019.

[26] Q. Yu and L. Sun, "LPClass: lightweight personalized sensor data classification in computational social systems," *IEEE Transactions on Computational Social Systems*, vol. 9, pp. 1–11, 2022.

[27] C. Luo, Z. Tan, G. Min, J. Gan, W. Shi, and Z. Tian, "A novel web attack detection system for Internet of things via ensemble classification," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5810–5818, 2021.

[28] J. Yan, Z. Du, J. Li, S. Yang, J. Li, and J. Li, "A threat intelligence analysis method based on feature weighting and BERT-BiGRU for industrial internet of things," *Security and Communication Networks*, vol. 2022, Article ID 7729456, 11 pages, 2022.

[29] X. R. Wang, J. Yang, Q. Y. Wang, and C. X. Su, "Threat intelligence relationship extraction based on distant supervision and reinforcement learning," *SEKE (2020)*, vol. 1, pp. 1–5, 2020.

[30] D. Diaz Lopez, M. Blanco Uribe, C. Santiago Cely et al., "Shielding IoT against cyber-attacks: an event-based approach using SIEM," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 3029638, 18 pages, 2018.

[31] Z. Liu, Z. Sun, J. Chen et al., "STIX-based network security knowledge graph ontology modeling method," in *Proceedings of the 2020 3rd International Conference on Geoinformatics and Data Analysis*, pp. 152–157, Marseille, France, 2020.

[32] L. Dandurand, "Cyber defense data exchange and collaboration infrastructure (CDXI)," May 2022, https://www.first.org/resources/papers/conference2010/dandurand-slides.pdf.

[33] J. S. Zhao, M. X. Song, X. Gao, and Q. M. Zhu, "Research on text representation in natural language processing," *Journal of Software*, vol. 33, pp. 102–128, 2022.

[34] Mitre, "Common vulnerabilities and exposure," May 2022, https://cve.mitre.org/.

[35] Mitre, "Common weakness enumeration," May 2022, https://cwe.mitre.org/.

[36] Mitre, "Common attack pattern enumeration and classification," May 2022, https://capec.mitre.org/.

[37] Z. Y. Liu, W. Y. Huang, Y. B. Zheng, and M. S. Sun, "Automatic keyphrase extraction via topic decomposition," in *Proceedings of the 2010 conference on empirical methods in*

natural language processing, pp. 366–376, Cambridge, Massachusetts, 2010.

[38] Z. Liu, W. Huang, Y. Zheng, and M. Sun, "An indicator of compromise extraction method based on deep learning," *Chinese Journal of Computers*, vol. 44, no. 5, pp. 882–896, 2021.

[39] G. Daniel and S. V. Thankachan, "Text indexing for regular expression matching," *Algorithms*, vol. 14, no. 5, p. 133, 2021.

[40] J. He, R. Yu, X. S. Wang, and L. N. Huang, "Validation of query expression based on regular expression," in *2011 International Conference on Computer Science and Service System (CSSS)*, pp. 1879–1882, Nanjing, 2011.

[41] J. Zhao, Q. B. Yan, J. X. Li, M. Shao, Z. He, and B. Li, "TIMiner: automatically extracting and analyzing categorized cyber threat intelligence from social data," *Computers & Security*, vol. 95, p. 101867, 2020.

[42] First, "Common vulnerability scoring system," May 2022, https://www.first.org/cvss/.

[43] S. Zhang, G. Bai, H. Li, P. Liu, M. Zhang, and S. Li, "Multi-source knowledge reasoning for data-driven IoT security," *Sensors*, vol. 21, no. 22, p. 7579, 2021.

[44] Infosec Consortium, "Inside report apt attacks on Indian cyber space," May 2022, https://app.box.com/s/a2zw9uye2hhofsc1me6yfj39u6gjalcq.

[45] Mitre, "ATT&CK matrix for enterprise," May 2022, https://attack.mitre.org/techniques/T1547/001/.

[46] Mitre, "ATT&CK matrix for enterprise," May 2022, https://attack.mitre.org/techniques/T1082/.

[47] Mitre, "ATT&CK matrix for enterprise," May 2022, https://attack.mitre.org/techniques/T1119/.

[48] K. Bandla and S. Castro, "APTnotes," May 2022, https://github.com/aptnotes/data.

[49] Mitre, "ATT&CK matrix for enterprise," May 2022, https://attack.mitre.org/techniques/T1129/.

[50] N. Tom and V. Claus, "Kill chain attack modelling for hidden channel attack scenarios in industrial control systems," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 11074–11080, 2020.