Automated Generation of Cyber Threat Intelligence

KAKUMARU Takahiro, TAKAHASHI Wataru, KATSUSE Riku, SIRACUSANO Giuseppe, SANVITO Davide, BIFULCO Roberto

Abstract

In order to identify cybersecurity risks at an early stage, NEC's intelligence analysts daily gather, accumulate, and analyze cybersecurity information. However, the scope of information to be gathered has expanded beyond cyberattacks, including information about political, economic, social, and technological trends. As a result, one of the challenges is how to appropriately narrow down sources of information and conduct an analysis while gathering information from a wider range of fields. NEC is working to automate the generation of cyber threat intelligence using generative AI for high-accuracy and rapid analyses.

This paper presents NEC's cyber threat intelligence initiatives and challenges as well as an extraction and summarization pipeline under development for cyber threat information and a search and analysis pipeline for cyber threat-related information.

Keywords

Generative AI, cyber threat intelligence, strategic intelligence, information extraction technology, synthetic analysis

1. Introduction

Cyber threat intelligence refers to the collection and analysis of information about cyber threats as a process as well as a deliverable[1)]. In general, there are three types of intelligence: tactical, operational, and strategic[2)]. Each of these types has a different position, purpose, and user.

In particular, strategic intelligence supports management teams in making decisions on cybersecurity risks by analyzing from a variety of angles not only cyber threats and cyberattacks surrounding the organization but also political, economic, social and technological factors in the external environment.

Information—mainly in regard to strategic intelligence—is written in a variety of formats including reports, blogs, articles, news, and social media, and different descriptions or expressions might be used by those sending out the information. Therefore, it is not easy to gather and organize the information. This is why intelligence analysts have been required to be highly knowledgeable and experienced. However, in the future, we will need to gather information even more broadly to find the desired information. To achieve this, we cannot rely solely on skilled analysts and we need to move away from overly relying on people to gather information.

In some cases, reports might be required within a few days or even a few hours after receiving a request for collection and analysis. To meet the expectations of the management team, speed is important to provide accurate and relevant information to enable decision making in a timely manner.

Against this background, we need to automate and save labor in the intelligence generation process to reduce the workload of analysts and to improve their analytical capabilities.

2. Strategic Intelligence Initiatives and Challenges

Users of strategic intelligence require information on what problems are currently occurring, what the situation is in other similar organizations, and what decisions are required to be made. Intelligence analysts gather necessary facts and use their knowledge to analyze them and formulate hypotheses, thereby meeting the expectations of management teams.

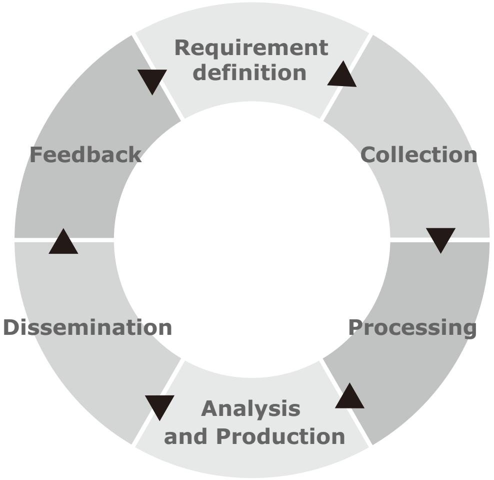


Fig. 1 Intelligence cycle.

The series of steps required to gather and analyze information upon request from decision makers and to generate the information necessary to take actions is called an intelligence cycle[3]. There are several variations of the intelligence cycle. The cycle described here includes the following steps: requirement definition, collection, processing, analysis/production, dissemination, and feedback (**Fig. 1**).

On the other hand, if you start to gather information from scratch only after receiving a request and then organize and categorize it, you will never complete it by the expected time if you consider that you must further analyze and appropriately organize the information that contributes to the decision making. It is thus important to gather and accumulate information on a daily basis and have it readily available for immediate use. In addition, it is also key to monitor information on a daily basis in order to be aware of emerging threats and changes in threats as soon as possible.

Therefore, NEC stores the information about cyber threats that is daily gathered and organized by intelligence analysts in a database. However, the scope of collection is no longer limited to cyber threat information but has expanded to include information regarding trends in politics, economy, society, and technology. Therefore, the challenge was to make the process of gathering, organization, and storage as labor-free as possible and to cover a wider range of fields in the information.

Furthermore, when investigating cyber threats, a vast amount of information sources must be read when searching for relevant information before collecting and analyzing the necessary facts. However, it can be difficult to find where the necessary information lays. For example, if the information matches at the keyword level but is not the specific information you are looking for, you cannot use it and end up wasting all that effort. Therefore, another challenge is to appropriately narrow down sources of information to precisely identify the ones required.

To solve these challenges, this paper presents a pipeline for extracting and summarizing cyber threat information, using the generative AI being developed by NEC, and an integrated pipeline for analyzing cyber threat-related information.

3. Policy on Achieving Automation of Cyber Threat Intelligence Generation Using Generative AI

Considering the importance of strategic intelligence, it is most important to solve the following challenges:

(1) Incorrect information may be included in the content generated by the generative AI model (hallucinations).

(2) The generative AI model may ignore important information that appears in long documents provided as input (long context).

As a solution to the first challenge, a generative AI model can be used as a reasoning tool. Specifically, the generative AI model is instructed to answer using solely the information explicitly provided in the input from outside sources. A verification process that would check the accuracy of answers given by the generative AI model is also introduced. This process includes a variety of approaches, including verification through external sources of information as well as self-review and reasoning steps that are conducted through interaction with the AI model.

As a solution to the second challenge, a preliminary processing step can be introduced before sending long texts to the generative AI model. In this phase, the original texts will be analyzed and only the contents that are highly relevant to the task at hand will be selected and extracted from the texts. This process of narrowing down the information can be adjusted upon request from the intelligence analysts and changed in accordance with the tasks to be automated.

4. Pipeline for Extracting and Summarizing Cyber Threat Information

4.1 Summary

When gathering and organizing information from a variety of non-structured sources of information such

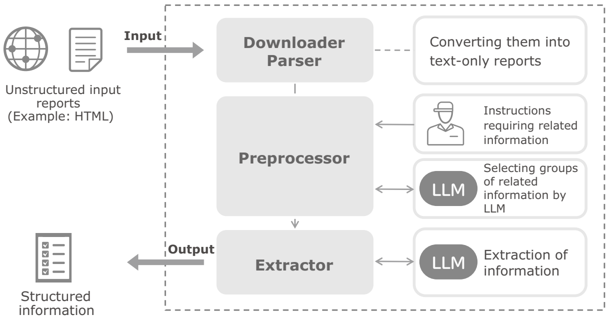

Input
Unstructured input reports (Example: HTML)
Downloader Parser
Converting them into text-only reports
Preprocessor
Instructions requiring related information
LLM
Selecting groups of related information by LLM
Output
Structured information
Extractor
LLM
Extraction of information

Fig. 2 Pipeline for extracting and summarizing cyber threat information.

as reports, blogs, articles, news, and social media, the intelligent analysts must manually read and make sense of a huge number of documents and then convert the information into a structured format that is appropriate and easily accessible in the database.

NEC's pipeline for extracting and summarizing cyber threat information aims to automate these tasks and make them more efficient. This pipeline consists of three elements (**Fig. 2**).

The downloader/parser converts input reports in a variety of unstructured formats such as HTML into a text-only format. For specific well-known sources of information, plugins to handle the formatting should be created in advance. These converted text documents are the only source of knowledge to be later used by the generative AI.

The preprocessor performs iterative filtering to identify relevant information. First, input texts are divided into short paragraphs, and each paragraph is adjusted to have a couple of sentences overlap with the previous paragraph to ensure that the context will not be lost. Then, generative AI is used to select relevant information from the newly formed paragraphs.

The specific types of relevant information required by the intelligence analysts can be easily configured with instructions that can be written in a natural language. For example, an instruction such as "Identify information about the economic impact of the attack" can be issued. Only the texts that are highly relevant to the task will be selected from the original texts and compiled into a group of relevant information.

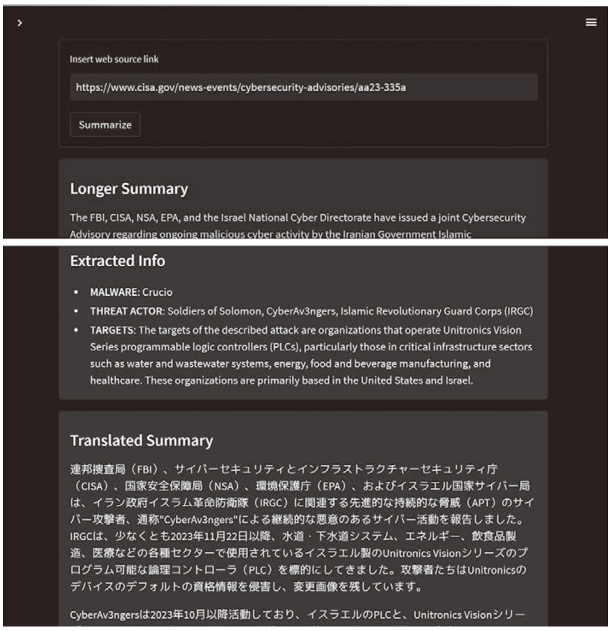
Finally, after receiving the information compiled in the previous step as input, the extractor stage extracts information by using generative AI and outputs it in a format that is consistent with the structured format to be used in the designated database. In this way, the extracted information is accumulated.

4.2 Features

The pipeline for extracting and summarizing cyber threat information has two important features. One feature is that it is possible to discern relevant information and focus the generative AI-assisted extraction on these details. In fact, when manually analyzing threat reports, the intelligence analyst must determine what to omit and what to include at the time of extracting information while maintaining the report's perspective. This determination generally involves considerations about the level of confidence as well as the level of detail of the information stated in the report. NEC's generative AI technology can easily be instructed to automatically perform analyses and select relevant texts like the analysts do. Another feature is that the pipeline can be used to swiftly adjust the format of the extracted information before outputting it in accordance with the required task or the requirements of the final recipient platform. By adjusting the instructions given to generative AI, this can be achieved both in the case of extracting specific information from texts and in the case of summarizing information.

4.3 Example of utilization

This pipeline is used to monitor a set of information sources and extract structured information. Specific information is extracted from texts and summarized,



Insert web source link

https://www.cisa.gov/news-events/cybersecurity-advisories/aa23-335a

Longer Summary

The FBI, CISA, NSA, EPA, and the Israel National Cyber Directorate have issued a joint Cybersecurity Advisory regarding ongoing malicious cyber activity by the Iranian Government Islamic

Extracted Info

- **MALWARE:** Crucio
- **THREAT ACTOR:** Soldiers of Solomon, CyberAv3ngers, Islamic Revolutionary Guard Corps (IRGC)
- **TARGETS:** The targets of the described attack are organizations that operate Unitronics Vision Series programmable logic controllers (PLCs), particularly those in critical infrastructure sectors such as water and wastewater systems, energy, food and beverage manufacturing, and healthcare. These organizations are primarily based in the United States and Israel.

Translated Summary

連邦捜査局（FBI）、サイバーセキュリティとインフラストラクチャーセキュリティ庁（CISA）、国家安全保障局（NSA）、環境保護庁（EPA）、およびイスラエル国家サイバー局は、イラン政府イスラム革命防衛隊（IRGC）に関連する先進的な持続的な脅威（APT）のサイバー攻撃者、通称"CyberAv3ngers"による継続的な悪意のあるサイバー活動を報告しました。IRGCは、少なくとも2023年11月22日以降、水道・下水道システム、エネルギー、飲食品製造、医療などの各種セクターで使用されているイスラエル製のUnitronics Visionシリーズのプログラム可能な論理コントローラ（PLC）を標的にしてきました。攻撃者たちはUnitronicsのデバイスのデフォルトの資格情報を侵害し、変更画像を残しています。

CyberAv3ngersは2023年10月以降活動しており、イスラエルのPLCと、Unitronics Visionシリー

Fig. 3 Example of utilization of information extraction and summarization.

and events are classified in accordance with the storage format of the database. This reduces the time required for a junior analyst to gather and summarize threat information by 50% from approximately two hours to one hour (**Fig. 3**).

5. Pipeline for Acquiring and Analyzing Cyber Threat-Related Information

5.1 Summary

The intelligence analyst's job is to process information from various sources of information, extract appropriate events, and correlate events and data from different documents and tools. The pipeline for acquiring and analyzing cyber threat-related information is designed to support the process used by the analyst and automate the operation as much as possible. This pipeline consists of three different elements (**Fig. 4**).

The search and retrieval module takes as input intelligence analysis queries expressed in natural language. For example, queries such as "Tell me about all attacks related to a specific threat actor" or "Is this file hash associated with threat actor X?" are accepted. Based on the first query, generative AI is utilized to perform retrieval augmented generation (RAG) over external sources of information.

A variety of threat knowledge bases can be used as external sources for this purpose. For example, proprietary threat databases, knowledge graphs that represent knowledge connections in a graph structure, vector stores indexed in vector form, and trustworthy publicly available information on the web are some types of such information that might be included.

These documents are input into the next module (relevance verification/extraction module) to verify if the searched documents are actually related to and relevant for the first query. If the relevance is confirmed, information from respective documents is extracted as a concise summary covering all the information related to the first query while keeping the original perspectives of the documents.

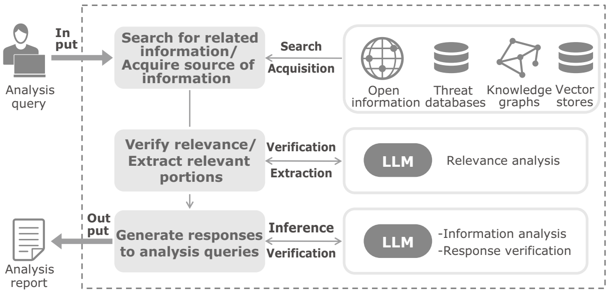


Fig. 4 Pipeline for acquiring and analyzing cyber threat-related information.

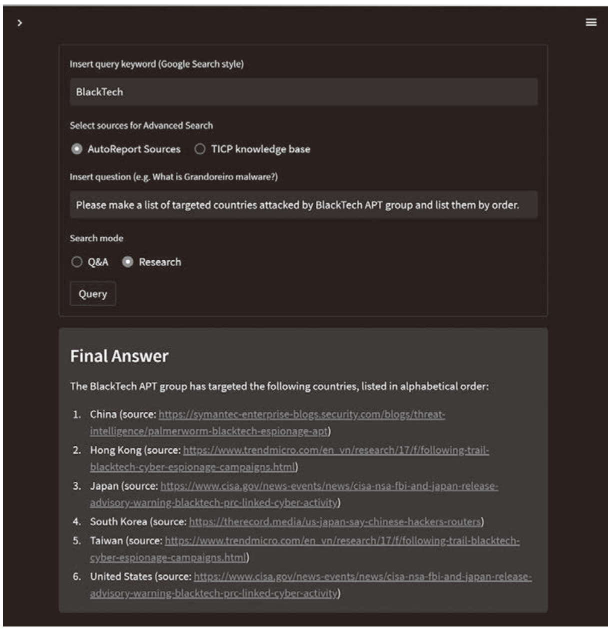
Finally, the relevance is checked, and the summarized documents are passed to the analysis module. In this module, generative AI is used as a reasoner that correlates information in the relevant summaries and generates an answer to the first query. External tools can be called in at this stage to further confirm the inferred analysis results. For example, services that analyze malware hashes and verification using threat databases are available.

5.2 Features

The main feature of this pipeline is to enable document analysis across highly reliable information sources by searching and processing a vast volume of information from a variety of threat databases and multiple documents. It identifies the relevance between the analysis query and documents and then summarizes the contents while maintaining the original perspective, making it possible to give a highly accurate answer. The accuracy will further improve by verifying the given answer with external tools or sources of information.

5.3 Example of utilization

Currently, the pipeline for searching and analyzing



Insert query keyword (Google Search style)

BlackTech

Select sources for Advanced Search

● AutoReport Sources ○ TICP knowledge base

Insert question (e.g. What is Grandoreiro malware?)

Please make a list of targeted countries attacked by BlackTech APT group and list them by order.

Search mode

○ Q&A ● Research

Final Answer

The BlackTech APT group has targeted the following countries, listed in alphabetical order:

1. China (source: https://symantec-enterprise-blogs.security.com/blogs/threat-intelligence/palmerworm-blacktech-espionage-apt)
2. Hong Kong (source: https://www.trendmicro.com/en_vn/research/17/f/following-trail-blacktech-cyber-espionage-campaigns.html)
3. Japan (source: https://www.cisa.gov/news-events/news/cisa-nsa-fbi-and-japan-release-advisory-warning-blacktech-prc-linked-cyber-activity)
4. South Korea (source: https://therecord.media/us-japan-say-chinese-hackers-routers)
5. Taiwan (source: https://www.trendmicro.com/en_vn/research/17/f/following-trail-blacktech-cyber-espionage-campaigns.html)
6. United States (source: https://www.cisa.gov/news-events/news/cisa-nsa-fbi-and-japan-release-advisory-warning-blacktech-prc-linked-cyber-activity)

Fig. 5 Example of utilization of information acquisition and analysis.

cyber threat related information has been internally verified in NEC. Intelligence analysts can interact with a system that incorporates generative AI through a web interface to acquire the automatically captured information and the analysis results (**Fig. 5**). Initial feedback indicates that it accelerates CTI operations while delivering accurate information.

6. Conclusion

In order to identify cybersecurity risks at an early stage, NEC's intelligence analysts gather, accumulate, and analyze cybersecurity information on a daily basis. NEC is working to automate and improve efficiency, using generative AI to further expand its sources of information and strengthen its analytical capability.

This paper presented an extraction and summarization pipeline for cyber threat information that can solve the challenges of hallucinations and long context for generative AI as well as a search and analysis pipeline for cyber threat-related information. We have verified some aspects and will continue to verify unproven areas. By promoting further incorporation of the analytical know-how developed by NEC, we plan to proceed with technological development to enable a more flexible and diverse threat analysis.

NEC will continue through its initiatives in cyber intelligence to contribute to solving social issues and creating social value in the form of fairness and efficiency.

References

1) ISO/IEC 27002:2022: Information security, cybersecurity and privacy protection - Information security controls.
https://www.iso.org/standard/75652.html

2) U.S. Joint Chiefs of Staff: Doctrine for Intelligence Support to Joint Operations, March 2000.
https://www.hsdl.org/c/abstract/?docid=3735

3) U.S. Director of National Intelligence: U.S. National Intelligence: An Overview, 2011.
https://www.dni.gov/files/documents/IC_Consumers_Guide_2011.pdf

Authors' Profiles

KAKUMARU Takahiro
Director and certified information systems security professional (CISSP)
Cyber Security Strategy Department

TAKAHASHI Wataru
Cyber Security Strategy Department

KATSUSE Riku
Cyber Security Strategy Department

SIRACUSANO Giuseppe
Principal Research Scientist
NEC Laboratories Europe

SANVITO Davide
Senior Research Scientist
NEC Laboratories Europe

BIFULCO Roberto
Senior Manager
NEC Laboratories Europe

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese

English

Vol.17 No.2 Special Issue on Revolutionizing Business Practices with Generative AI

– Advancing the Societal Adoption of AI with the Support of Generative AI Technologies

Orchestrating a brighter world
NEC
NEC Technical Journal
Vol.17 No.2
June 2024
Special Issue on Revolutionizing Business Practices with Generative AI
Advancing the Societal Adoption of AI with the Support of Generative AI Technologies

Vol.17 No.2

June 2024

Special Issue TOP

Papers for Special Issue

Market Application of Rapidly Spreading Generative AI

Fundamental Technologies that Enhance the Potential of Generative AI

For AI Technology to Penetrate Society

NEC Information