Multi-labeling with topic models for searching security information

Osada, Yuki ; Nagasawa, Ryusei ; Shiraishi, Yoshiaki ; Takita, Makoto ; Furumoto, Keisuke ; Takahashi, Takeshi ; Mohri, Masami ; Morii,…

(Citation)
Annals of Telecommunications,77(11):777-788

(Issue Date)
2022-12

(Resource Type)
journal article

(Version)
Version of Record

(URL)
https://hdl.handle.net/20.500.14094/0100477800

Annals of Telecommunications (2022) 77:777–788
https://doi.org/10.1007/s12243-022-00928-5

Multi-labeling with topic models for searching security information

Yuki Osada[1] · Ryusei Nagasawa[1] · Yoshiaki Shiraishi[1] · Makoto Takita[2] · Keisuke Furumoto[3] · Takeshi Takahashi[3] · Masami Mohri[4] · Masakatu Morii[1]

Received: 14 April 2021 / Accepted: 6 October 2022 / Published online: 26 October 2022

Abstract
Security information such as threat information and vulnerability information are utilized to analyze cyberattacks. If specific keywords such as the name of malware related to the event to be analyzed are known in advance, it is possible to obtain information using typical search engines. However, when a security operator cannot recall appropriate keywords related to the event to be analyzed, or when a commonly recognized identifier does not exist, a general search engine cannot be expected to produce useful results. In this paper, we propose a method using topic models and outlier detection to generate multi-labels for search, with the goal of constructing a search engine that can present relevant security information even in such situations. In addition, this paper discusses the application of the proposed method to 2386 security reports issued from 2017 to 2019 to demonstrate that the labeling can be focused on specific topics.

Keywords Multi-labeling · Security reports · Threat intelligence · Topic models

1 Introduction

Cyberattacks targeting organizations are becoming more sophisticated and complex as the Internet becomes another form of infrastructure. Organizations must take pre-incident measures before an attack occurs, as well as take prompt post-incident countermeasures following an attack. Security reports that summarize the causes of incidents, attack methods, etc. are useful for precautionary measures and post-incident response.

The number of security reports and other security information published regularly by security vendors increases every day. However, there is no standard for assigning labels for retrieval, and such labels therefore may vary based on the issuer. Some documents are not labeled at all. Against this backdrop, it is not possible to use the labels assigned to documents for cross-searching security information in multiple information sources. Therefore, for security operators to retrieve the desired security information, a centralized label according to the content of the document is required. Multi-labeling is an appropriate technique, as there are several important words and topics that can be used as keywords in a single document.

Topic models and keyword extraction methods are used for multi-labeling. Topic models can analyze latent topics that do not appear in documents based on the co-occurrence of words in the documents. A typical example of a topic model is Latent Dirichlet Allocation (LDA) [1], which was proposed by Blei et al. There are three main types of keyword extraction methods: statistical-based, graph-based, and machine learning-based. Typical approaches for each method include RAKE [2], TextRank [3], and KEA [4]. The objective of this research was to attach multi-labels to documents that allow security operators to collect a variety of related information associated with the retrieved matter. We believe that the topic model is the best method to achieve this goal. Keyword extraction methods have limited accuracy in extracting keywords and can select words incorrectly as labels in the document. Therefore, only names of malware or attack campaigns in the document can be extracted, and no useful multi-label can be attached to the documents.

✉ Yoshiaki Shiraishi
zenmei@port.kobe-u.ac.jp

[1] Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan

[2] University of Hyogo, 8-2-1 Gakuennishi-machi, Nishi-ku, Kobe, Hyogo 651-2197, Japan

[3] National Institute of Information & Communications Technology, 4-2-1, Nukui-Kitamachi, Koganei, Tokyo 184-8795, Japan

[4] Kindai University, 3-4-1 Kowakae, Higashi-Osaka 577-8501, Japan

However, with LDA, it is possible to label keywords that are not in the document.

We propose a method for multi-labeling security reports using topic models and a method for improving the generalization performance of the models. First, we utilize an outlier detection algorithm to identify documents whose contents are considered different from others as outlier documents, and then build a topic model by excluding outlier documents. Next, the constructed topic model is used to dimensionally compress the document vector of security reports, and the low-dimensional vector is clustered to collect documents with similar potential topics. Then, multi-labels are assigned to the documents in each cluster based on the parameters of the topic model. Here, there are numerous security reports that contain terms related to multiple topics (henceforth, referred to as "summary documents") for monthly reports and alerts. When summary documents are labeled, the label quality is low to the extent that terms that do not fit the contents of the documents or terms with broad meanings are selected. If we improve the quality of labels for such documents, the quality of document retrieval results will be improved.

This paper discusses the goal of improving the quality of security information retrieval results by finding summary documents and further vectorizing the documents by the appropriate topic. Specifically, to identify summary documents, we first cluster them by LDA document vectors to find clusters. Then, we propose a method for building, vectorizing, and labeling the LDA model again using the documents belonging to the cluster. By applying the proposed method to 2,386 security reports issued by eight security vendors from 2017 to 2019, we confirmed that the labeling can be more focused on specific topics.

2 Multi-labeling using topic models and outlier detection

2.1 Latent Dirichlet Allocation (LDA)

A topic model is a method for latent semantic analysis which can analyze potential topics based on the co-occurrence of words in a document. LDA, a type of topic model, assumes that there are multiple latent topics in each document, and analyzes topics that do not appear in the document by inferring topics from the co-occurrence information of words in the document.

By inputting a set of documents for training and specifying the number of topics, LDA is trained, and topics are generated. When a set of documents represented by a high-dimensional vector is input to LDA, each document is vectorized by the probability of belonging to each topic, Θ. Additionally, each topic is vectorized by the probability distribution Φ with respect to the set of words that constitute the topic. Because the document vector is converted to a vector with the dimension of the number of topics, LDA can be used as a method of dimensional compression.

2.2 Outlier detection algorithm

In the following, the antonym of an outlier value is referred to as a "normal value." One-Class SVM [5] is an extension algorithm of a support vector machine, which is a one-class classification method. The dataset is mapped from the input space to the feature space using a nonlinear kernel function, and outlier detection is performed by drawing a discriminative boundary that separates normal values from outliers.

2.3 Improving topic model performance deficiencies through outlier detection

Generalization performance can be improved by applying the outlier detection method to a set of document vectors constructed by LDA with a certain document set, as well as by constructing a model by LDA with only normal-valued documents. Then, the document set can be clustered to form clusters amenable to multi-labeling. This is because outlier detection can be applied to form more specialized topics, and by vectorizing them with the LDA model, clusters specialized for each topic can be formed. However, if a document contains terms related to multiple topics, the vector will belong to multiple topics. When such documents are labeled based on the probability distribution Φ of the set of words that make up the topic, the quality would be low, as the labels are not semantically consistent, and terms with broad meanings or terms that would not fit the content of the document are chosen. That is, we have a challenge towards improving label quality with the goal of constructing a search engine that can present relevant security information.

3 Proposed methodology

We improved the quality of the labels by training LDA again on such a set of documents and vectorizing them with respect to the task of assigning appropriate labels based on the probability distribution Φ of the set of words constituting a topic to documents in which there are multiple terms related to a topic in a single document. The flow of the proposed method is shown in Fig. 1, and the specific process is described below.

Step 1 Preprocessing

As a preprocessing step for security reports, we extract the title, text, and captions of figures and tables from each

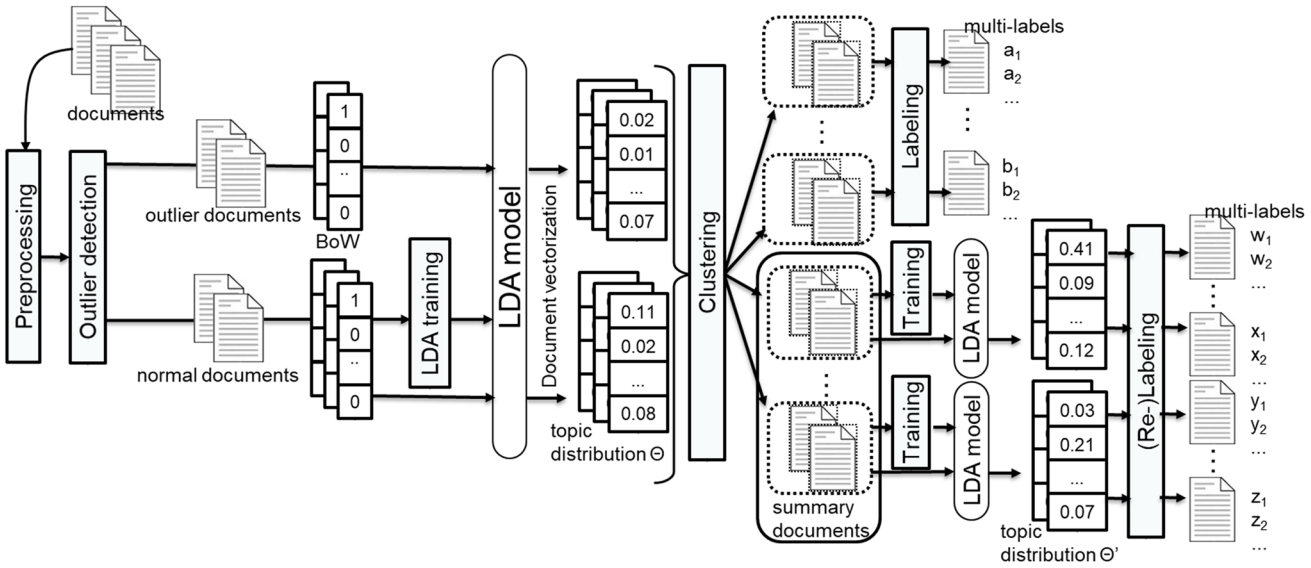


Fig. 1 Flow of the proposed method

report and split them into words. In addition, as technical terms used in the security industry such as "browser base" and "mobile device" appear in security reports, we extract such terms by using $N = 2$ of N-grams, which divides a document into N consecutive words to make compound words. In addition to stop words such as prepositions, articles, and pronouns, this also removes descriptions that are not directly related to the content of the text, such as malware hash values, URLs, and IP addresses of C&C servers. The bag-of-words (BoW) technique, which vectors documents based on the frequency of occurrence of words, does so based on the frequency of occurrence of words and compound words.

Step 2 Outlier detection

The parameters of the algorithm are set to default and One-Class SVM is applied to the document vector. The outlier documents are called "outlier documents" and the other documents are called "normal documents."

Step 3 LDA training

When constructing an LDA, the user must specify the number of topics in advance based on the number of documents and their content. In this study, we used the Arun_2010 [6] and Coherence_mimno_2011 [7] evaluation functions available in the Python tmtoolkit package [8] to derive the number of topics, as shown in the following steps.

Step 3-1 Calculate the values of Arun_2010 and Coherence_mimno_2011 for some candidate values of the number of topics given in advance.

Step 3-2 Normalize the evaluation values of Arun_2010 and Coherence_mimno _2011 to unify the data scale.

Step 3-3 The minimum value is the optimal value for Arun_2010, and the maximum value is the optimal value for Coherence_mimno_2011. The value of Arun_2010 is reversed, and the maximum value is the optimal value.

Step 3-4 Calculate the average value of the evaluation values for each topic number, and the number of topics with the maximum value is set as the optimal number of topics k.

We input the number of topics and normal documents to construct an LDA model.

Step 4 Document vectorization

The entire document is input into the constructed LDA model. Each document is dimensionally compressed into a topic distribution $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ based on the number of topics k obtained in Step 3. Each topic is also represented by a word distribution $\Phi = (\varphi_1, \varphi_2, \ldots, \varphi_n)$ based on the total number of words and compound words n obtained in Step 1.

Step 5 Assigning labels to documents

Each document is assigned a multi-label of compound words using Θ and Φ, based on the method as follows:

Step 5-1 In document d, probabilistically select a topic using vector θ_d.

Step 5-2 Select a term probabilistically by using the word distribution Φ_k in the selected topic k.

Step 5-3 Repeat Steps 5–1 and 5–2 1000 times for each document d and label d with the top 50 terms selected most frequently.

Step 6 Extraction of summary documents

We extract only summary documents and re-label them. First, we cluster all the documents by the document vector Θ, and set the average of the document vectors in each cluster as the center vector of the cluster. The number of clusters is set to the same value as the number of topics k. The following shows how to calculate the center vector. Let N be the number of documents in a cluster g. The vector of documents in a cluster is represented by $\Theta = (\theta_1, \theta_2, ..., \theta_k)$. By summing up the components of each document vector and dividing by the number of documents in N, we can obtain the center vector c_j of cluster g_j.

For $i = 1, 2, 3, \ldots, N$ do

$$\boldsymbol{v}_j += \left(\theta_{i_1}, \theta_{i_2}, ..., \theta_{i_k}\right)$$

End

$$\boldsymbol{c}_j = 1/N \cdot \boldsymbol{v}_j$$

Repeat this for all clusters and get the center vector $\boldsymbol{c}_j$ for each cluster $\boldsymbol{g}_j$.

Here, the vector of documents that deal with multiple topics in one document is thought to be likely to belong to a broad range encompassing multiple topics, and the difference between the components of the vector is small. Accordingly, we identify clusters where the difference between the top two components of the central vector of each cluster is less than a threshold value. The documents in the clusters are believed to be documents dealing with multiple topics, and are referred to as *summary documents*.

Step 7 Re-labeling summary documents

For each cluster obtained in Step 6, Steps 3–5 are repeated to obtain the document vector Θ′ and the word affiliation probability Φ′. Then, using Θ′ and Φ′, the multi-label of each summary document similarly can be obtained as in Step 5.

4 Evaluation of summary document labels

4.1 Experimental environment and dataset

The dataset used in our experiments consisted of 2386 security reports issued by eight different security vendors (Trendmicro [9], Cisco [10], Symantec [11], Barracuda [12], Druva [13], FireEye [14], Arbor [15], Palo Alto [16]) from January 1, 2017 to December 31, 2019. Python and Pandas were used to manage the data. NumPy was used to perform numerical computations on the data. The following steps were implemented to realize the proposed method:

Step 1, we used gensim.corpora.dictionary.Dictionary to convert the preprocessed security report into BoW vector. In Step 2, we used sklearn.svm.OneClassSVM to perform outlier detection. In Step 3, we created the LDA model using guidedlda.GuidedLDA [17]. For training the LDA model, we used tmtoolkit.topicmod.tm_lda.evaluate_topic_models to determine the number of topics. To determine the number of topics, we used sklearn.preprocessing.MinMaxScaler for normalization, to unify the data scale of the two evaluations, Arun_2010 and Coherence_mimno_2011. In Step 6, we used sklearn.cluster.KMeans to perform clustering based on the document vector θ. We also calculated the central vector using numpy.linalg.norm and numpy.dot.

4.2 Evaluation of document vectors and labels

We confirmed that the re-labeled result of the summary document in Step 7 is more content-specific than the label of the summary document in Step 5. Specifically, we compared the document vectors θ and θ′ of the summary documents and compared the labels based on the vectors and confirmed that the labels are more appropriate to the content.

Nine hundred fifty-four outliers were identified by One-Class SVM, and the LDA was trained to vectorize and cluster all the documents. There were 12 topics and clusters during Step 3 and Step 6. The central vectors for each cluster are shown in Fig. 2. Summary documents are the documents in the clusters numbered 0, 1, 7, and 10.

We compared the central vector of each cluster with the vectors of summary documents and evaluated the labels of the three documents with the highest cosine similarity. The vectors of the documents to be evaluated in each cluster are shown in Figs. 3, 4, 5 and 6. Figure 3a shows that the document vector Θ of document number 1556 belongs to multiple topics. However, Fig. 3b shows that the document vector Θ′ is specialized for topics 3 and 7. In the same way, we can see that the document vectors Θ in the lefthand side of all Figs. 4, 5 and 6 are specialized for certain topics in the document vector Θ′ on the righthand side. By labeling with Θ′, we can restrict the labeling to a specific topic. There was no correlation between the topic numbers of Θ and Θ′.

The label for document number 1556 is shown in Table 1. The content of the document is related to the "Triton" malware that affected industrial control systems in the Middle East. According to Table 1a, the most characteristic labels by document vector Θ are "email, machine learning, cloud, President Trump, wire transfer, data protection, Congress,

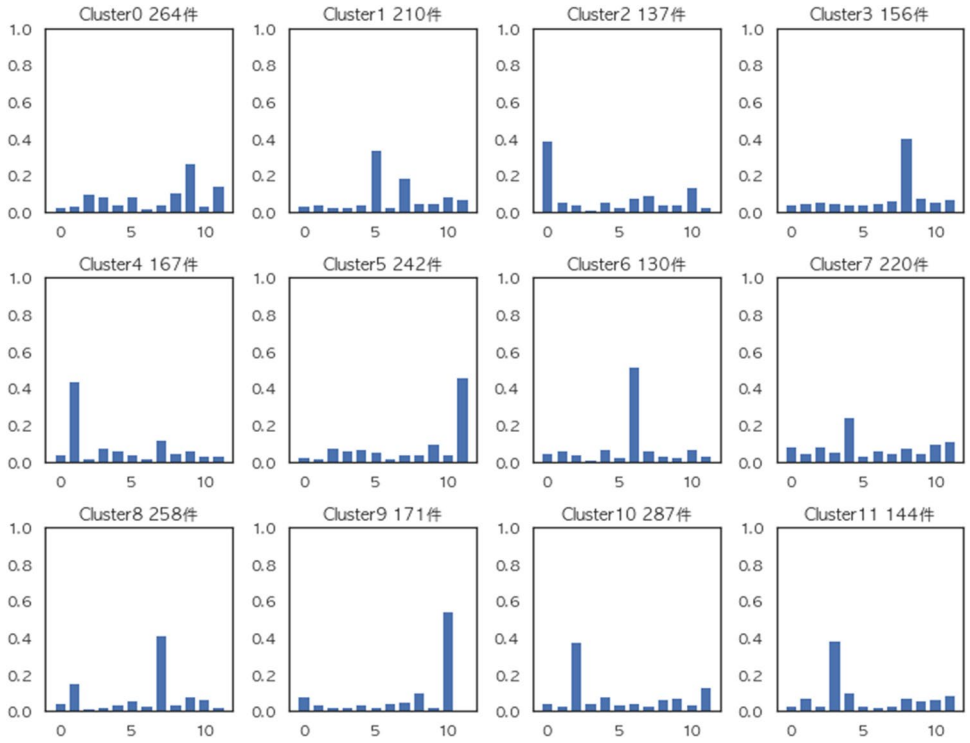
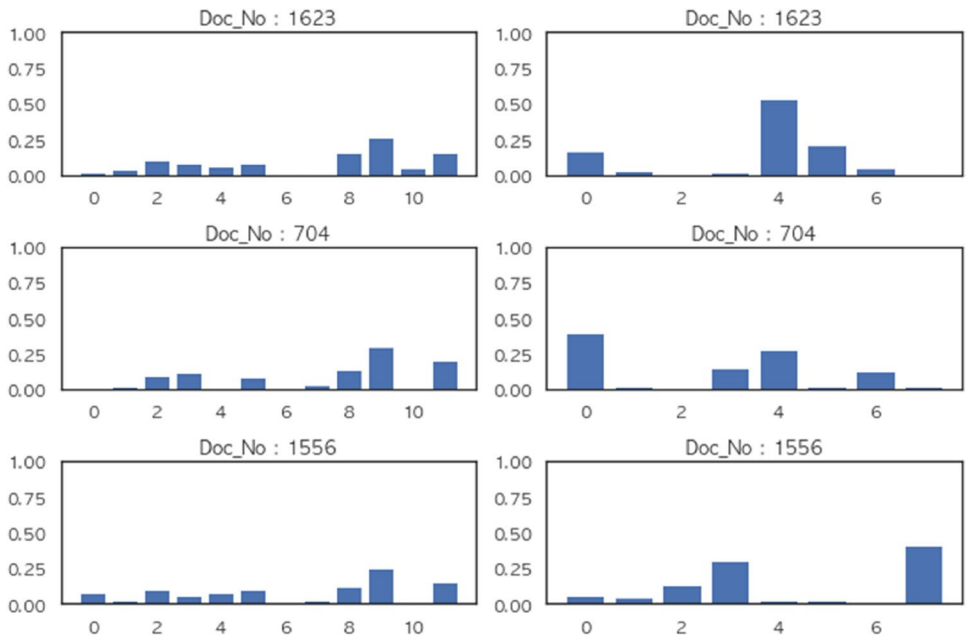


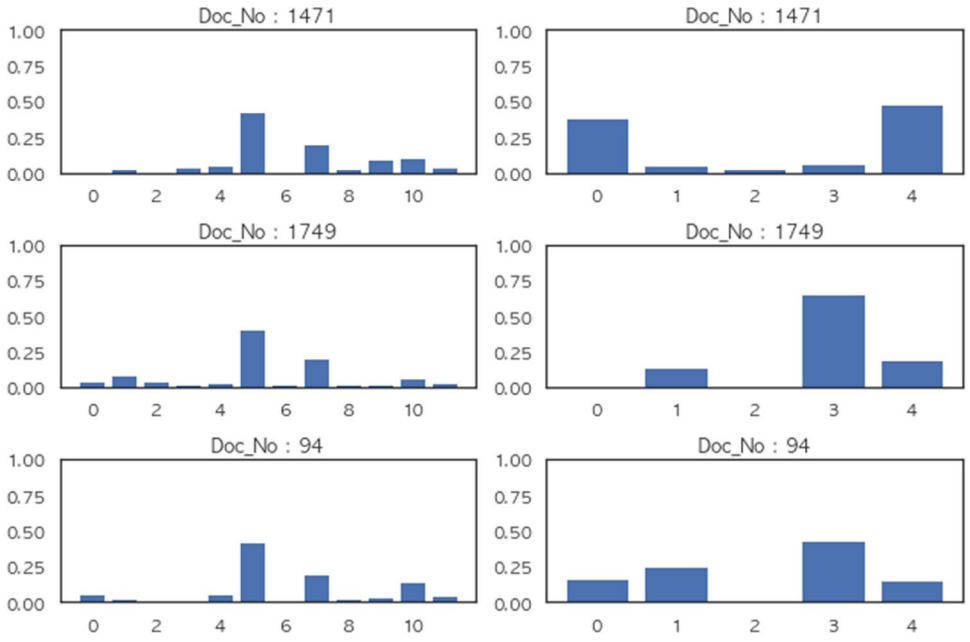
Fig. 2 Central vectors of each cluster



(a) Document vector Θ (b) Document vector Θ'

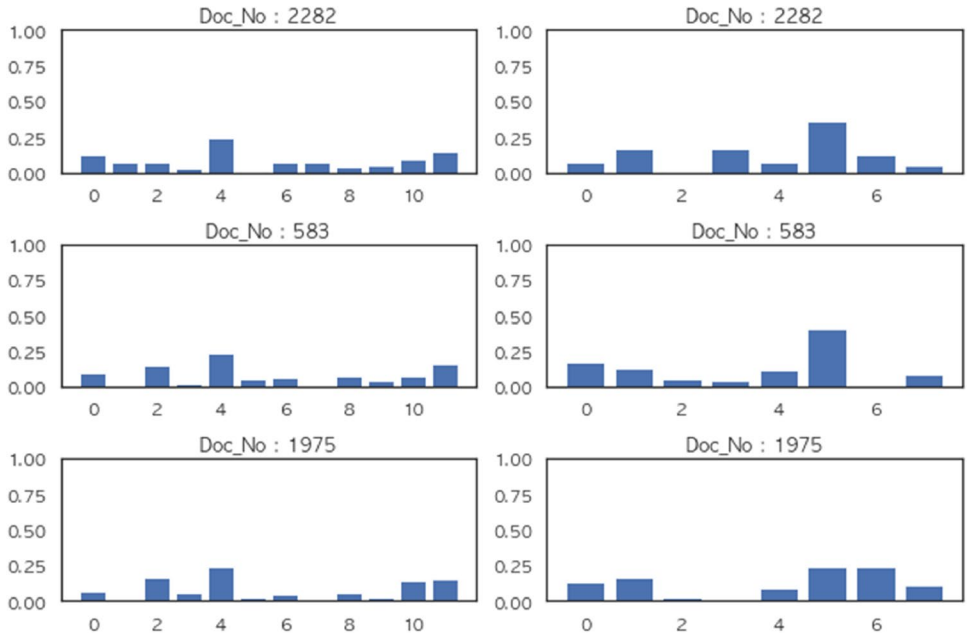
Fig. 3 Document vector in cluster 0

law enforcement, security products, supply chain, patch, IoT, remote control, mobile apps, web apps, OS/operating system, and data breach" and it is apparent that there are labels for a wide range of topics. Meanwhile, the labels assigned by the document vector Θ′ were more consistent with the content than (b): "IP address, source code file, cyber espionage, config file, targeted attack, phishing, APT attack, remote control, South Asia, malware framework, credential theft, DNS, lifeline, financial institutions, custom malware, OS/operating system, and MAC address," which are similar to the contents of other documents. We confirmed that the proposed method improves the quality of labels for summary documents.



(a) Document vector Θ (b) Document vector Θ'

Fig. 4 Document vector in cluster 1



(a) Document vector Θ (b) Document vector Θ'

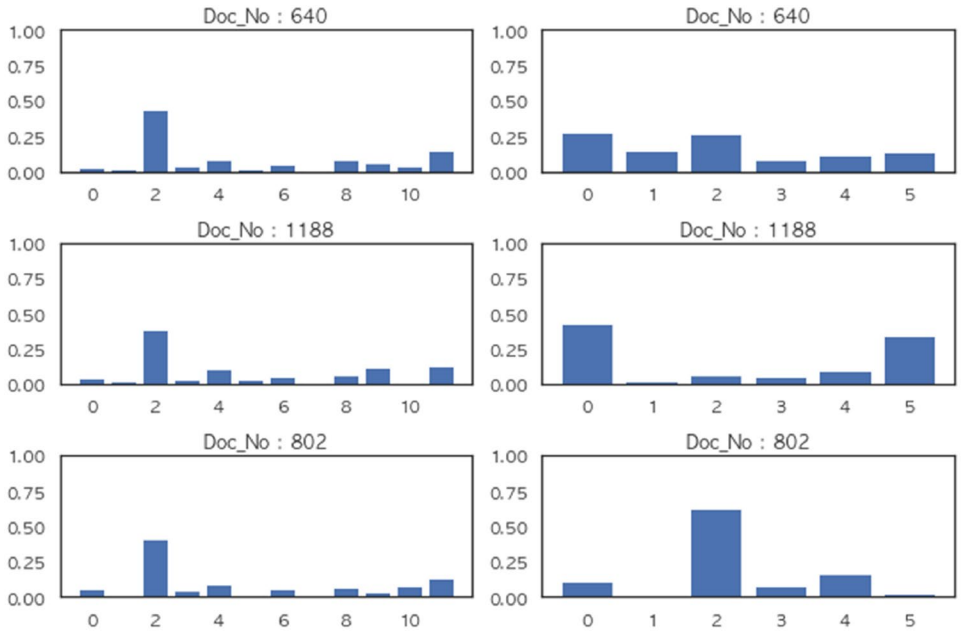
Fig. 5 Document vector in cluster 7

5 Case study

First, in subsection 5.1, we discuss if the labels generated by the proposed method represent the topics of the documents well, and if there is a label that adequately represents the topic of a document, it can be used as the next search query for security information retrieval. Next, in Sect. 5.2, we summarize and discuss the results of Sect. 5.1.

5.1 Topic analysis of similar documents focusing on reference documents

We focused on seven characteristic keywords, "Emotet," "Wannacry," "Shamoon," "Mirai," "Triton," "Samsam," and "Mobile Malware," and investigated the types of documents that can be obtained using them. In the following paragraphs, we refer to documents that contain these keywords as "reference documents."



(a) Document vector Θ (b) Document vector Θ'

Fig. 6 Document vector in cluster 10

Table 1 Labels for document number 1556

(a) Labels by document vector Θ

Labels for document number 1556 by vector Θ

email security, threat actor, cyber security, threat landscape, machine learning, cloud ready, ll help, trump administration, endpoint security, security pro, incident response, ransomware attack, data protection, re able, email threat, data breach, wire transfer, gain access, few day, threat intel, cyber threats, past few day, law enforcement, black hat, zero day, federal government, cyber criminal, threat intelligence, few year, security team, security product, endpoint protection, supply chain, security news, cloud security, security research, email attack, re available, internet of thing, hybrid environment, virtual patch, remote access, ll continue, mobile app, social engineering, web application, security advisory, operating system, data breaches, security control

(b) Labels by document vector Θ′

Labels for document number 1556 by vector Θ′

threat actor, ip address, registry key, source code, scheduled task, file name, cyber espionage, configuration file, targeted attack, phishing email, nocase ascii, re able, persistent threat, home page, same c, persistence mechanism, remote access, malicious act, cab file, user account, threat group, espionage group, threat intel, southeast asia, victim organization, malware framework, compile time, credential theft, carbanak backdoor, new malware, pdb path, open source, reverse engineering, victim environment, file system, dns record, log file, critical infrastructure, reverse engineer, custom tool, shell script, financial institution, malicious actor, custom malware, red team, operating system, available tool, config file, build tool, mac address

First, all documents were clustered by the document vector Θ. Each group of summary documents was clustered using the document vector Θ′ and then classified into subclusters. Next, we identified the cluster with the highest concentration of reference documents for each keyword, which is the set of documents most related to the keyword. Based on the results, we then visually examined the topics of the documents in the cluster and classify whether each topic is a superordinate concept or an equivocal concept with respect to the keyword.

The first case of "Emotet" is as follows: Emotet was first observed in 2014 and was initially reported as a malware that obtains credentials. However, Emotet has recently been used in more sophisticated targeted attacks, such as impersonating real organizations or using the contents of emails used in actual business activities to infiltrate PCs, steal personal information, and infect PCs with other malware.

Using the method as in Step 5.1, we analyzed the clusters that included reference documents for "Emotet." There were 21 reference documents for "Emotet" among all documents, and 8 documents each were clustered in cluster 10 and cluster 0. We divided clusters 0 and 10 into subclusters and found that there were 3 documents in clusters 0_2 and 0_7, and 4 and 2 documents in clusters 10_0 and 10_4, respectively. Cluster 10_0, which had the most referenced documents, is the same cluster as "Mirai." However, the documents in this cluster alone did not provide enough information about Emotet. Therefore, we analyzed the topics corresponding to clusters 0_2, 0_7, and 10_4.

Table 2 shows the results of the topic analysis. From cluster 10_4 in (a), we obtained information about finance related to Emotet. From cluster 0_2 in (b), we obtained information about spear phishing, a means of infection used by Emotet, as well as security information about Word, Excel, and RTF files used by Microsoft Office. From cluster

Table 2 Cluster topics for "Emotet"

(a) Cluster 10_4 topics	
Cluster 10_4 (12 items)	
Upper-level concept (number of documents)	Ransomware (2), DDoS(8), IoT(9), Mirai(8)
Equivalent concept (number of documents)	Emotet (1), MongoDB (1), bitcoin (2), AWS (1), DeOS (1), PDoS (1), Reaper (2), Adobe Flash Player (1), MAC (1), Dreambot (1), IcedID (1), bank-targeting Trojan horse (1), Hakai (1), OMG (1), Satori,JenX,Hajime (4), medical device (1), telnet (2), brute force attack (2), Realtek (3), Android (1), Linux (1), DemonBot (1), Hadoop YARN (1)
(b) Cluster 0_2 topics	
Cluster 0_2 (36 items)	
Upper-level concept (number of documents)	Ransomware (5), spearfishing (9)
Equivalent concept (number of documents)	Emotet (1), Trojan.Pandex (1), APT33 (1), FIN6 (1), APT41 (1), Trickbot (1), Rig (1), Ransom. Cerber (1), Raosom.Cry (1), Trojan.Wortrilk (1), APT29 (1), Fileless malware (3), Zero-day attacks (5), Microsoft OfficeRTF document (9), FINSPY (3), LATENTBOT (1), CARBANAK (1), Wannacry (6), SMB (3), EternalBlue (4), cryptocurrency miner (1), Backdoor.Nitol (1), credential information (1), CobaltStrike (3), Petya/NotPetya (2), LNK (5), Mimikatz (1), Excel (1), APT34 (2), SANNY (2), financial Trojan horse (5), Metamorfo (1), Shamoon (1), APT10 (1), Ryuk (2)
(c) Cluster 0_7 topics	
Cluster 0_7 (32 items)	
Upper-level concept (number of documents)	C&C (6), Win32 (5),
Equivalent concept (number of documents)	Gss3.x (1), APT28 (1), Greenbug (1), Shamoon (1), Disttrack (1), Ryuk (1), Ismdoor (1), FIN7 (2), SDB (1), CARBANAK (1), APT28 (1), Turia (1), APR41 (1), RediModiUpd (1), MatrixBanker (1), xdata (1), LockPoS (1), Fokibot (1), mimikatz (1), financial Trojan horse (4), Rig (2), Grobios (1), monero (1), NSISLoader (1), Delphi (1), COM object (2), HAWKBALL (1), Emotet(1), Trickbot (1),

0_7 in (c), we obtained information on a chain attack using Emotet, the bank-targeting ransomware Trickbot, and the ransomware Ryuk. In addition, we obtained information on several malware similar to Emotet. Thus, even if the number of documents containing a keyword was small, security information on the keyword could be obtained by browsing related documents in multiple clusters.

The results of the remaining keywords are as follows: There were 159 reference documents for the keyword "Wannacry" among all the documents, and 61 were collected in cluster 10. Similarly, 10 of 18 documents for "Shamoon" were in cluster 5, 28 of 57 documents for "Mirai" were in cluster 10, 7 of 9 documents for "Triton" were in cluster 0, 13 of 19 documents for "Samsam" were in cluster 10, and 14 of 30 documents for "Mobile Malware" were in cluster 11. Here, "Triton" was found in cluster 0. As cluster 0, where "Triton" was gathered, and cluster 10, where "Wannacry," "Mirai," and "Samsam" were gathered, are summary documents, the documents in each cluster were divided into subclusters. Among the sub-clusters of cluster 10, the clusters with the highest concentration of reference documents were "Wannacry" (3), "Mirai" (0), and "Samsam" (1). Cluster 5, which was related to "Shamoon," was not determined to be a set of summary documents. However, the result of the topic analysis showed that the topic was too broad to be used in the analysis. As such, we re-vectored and divided the cluster into sub-clusters to investigate the topic.

We surveyed the topics of the clusters with the highest concentration of reference documents and their subclusters and classified them into superordinate and equivocal concepts, the results of which are shown in Tables 3, 4, 5, 6, 7 and 8.

Table 3 shows the results of the topic survey on "Wannacry." First, although cluster 10, which was a cluster of summary documents, can be used to obtain information on a malware like Wannacry from the results of classification based on ranking concept, a number of different topics exist, such as IoT, financial institutions, and medical institutions. By contrast, (b) sub-cluster 3 focused on Wannacry. Documents related to the Lazarus and EternalBlue attacks, which were carried out by the Wannacry cyberattacker group, were present. In addition, there were documents related to Lazarus' targeted attacks on financial institutions and ATMs, as well as documents related to the EternalBlue attack, such as Petya and BadRabbit, confirming the existence of a group of documents related to the keyword "Wannacry."

Table 4 shows the results of the topic survey on "Mirai." (a) Cluster 10 is a set of summary documents and contains a variety of topics. By contrast, (b) sub-cluster 0 is a cluster focused on Mirai and IoT. We obtained information related to the IoT, such as the use of Telnet and SSH vulnerabilities to infect IoT devices, the discovery of Mirai variants with various exploits, and the increase in attacks

Table 3 Cluster topics for "Wannacry"

(a) Cluster 10 topics	
Cluster 10	
Upper-level concept (number of documents)	Ransomware (75), IoT Products (11), Banking (7), Medical Devices (5), CPU (6), Software (5), Hardware (6) Vulnerabilities, Healthcare (5)
Equivalent concept (number of documents)	Mirai (13), Dark Web (3), Phishing Emails (41), Targeted Attacks (6), Wannacry (34), Satan (1), JenX (1), Banking Trojan (13), Petya (21), NotPetya (10), Eternal Blue (16), ZDI (16), Bitcoin (4), VMWare (9), Sage (1), Cryptocurrency (7), Hajime (2), Satori(1), BadRabbit (4), ATM (5), Cryptolocker (5), Samsam (4), Shadow Broker (6), APT28 (2), Jigsaw (1), Lazarus (4), kirk (1), Spectre•Meltdown (5)
(b) Cluster 3 topics subordinate to cluster 10	
Cluster 10_3	
Upper-level concept (number of documents)	Wannacry (22), Targeted Attacks (4)
Equivalent concept (number of documents)	Industrial Control Systems (1), Exploit Kits (4), Business Email Compromise (BEC)(2), Pawn-Storm (2), Rig (1), Buckeye (1), Mining (2), Petya/NotPetya (15), Finance (8), Ransomware (7), IoT (1), EternalBlue (7), ATM Attacks (4), Lazurus (4), Attacks on Financial Institutions (3), BadRabbit (2), Cryptocurrency (2), Attacks on Health Care Industry (2), Adylkuzz Cryptocurrency Miner (1), Attacks on Insurance Industry (1), samsam (1)

Table 4 Cluster topics for "Mirai"

(a) Cluster 10_4 topics	
Cluster 10_0	
Upper-level concept (number of documents)	IoT (23), Mirai (20)
Equivalent concept (number of documents)	IIoT (6), Reaper (1), telnet (2), 5G (2), Hidden-Wasp (1), Mobile security (2), Supply chain (2), VPNfilter (2), Hakai DDoS bot (1), Carbanak (1), coin mining (3), Shodan (3), PDoS (1), NAS device vulnerabilities (5), Exploit Kits (5), Dyn (3), Wannacry (5), DDoS (13), Ransomware (5), SSH (2), Shadow Brokers (2), Petya/NotPetya (3), ML (1), Phishing (2), Satori/JenX/OMG/Wicked (1), Targeted Attack (2)
(b) Cluster 0 topics subordinate to cluster 10	
Cluster 10_0	
Upper-level concept (number of documents)	IoT (23), Mirai (20)
Equivalent concept (number of documents)	IIoT (6), Reaper (1), telnet (2), 5G (2), Hidden-Wasp (1), Mobile security (2), Supply chain (2), VPNfilter (2), Hakai DDoS bot (1), Carbanak (1), coin mining (3), Shodan (3), PDoS (1), NAS device vulnerabilities (5), Exploit Kits (5), Dyn (3), Wannacry (5), DDoS (13), Ransomware (5), SSH (2), Shadow Brokers (2), Petya/NotPetya (3), ML (1), Phishing (2), Satori/JenX/OMG/Wicked (1), Targeted Attack (2)

targeting illicit cryptocoin mining. We also obtained information on Gafgyt and its variant Hakai, which are similar malware to Mirai.

Table 5 shows the results of the topical survey on "Samsam." (b) Sub-cluster 1 obtained documents about Wannacry, a ransomware program similar to Samsam, and information about security in the healthcare industry, which was the primary target of Samsam. We also obtained information indicating that "ransomware-as-a-service" applications are making ransomware-based attacks increasingly easy to execute.

Table 6 shows the results of the topical survey on "Triton." (a) Cluster 0 provides information on cyberattacker groups such as APT28 and information on IoT security. Contrarily, (b) sub-cluster 6 provides information about Stuxnet, a malware that targeted Iran in 2010, and Industroyer, a malware that targeted Ukraine in 2016. These are malware programs similar to Triton that target industrial control systems.

Table 5 Cluster topics for "Samsam"

(a) Cluster 10 topics	
Cluster 10	
Upper-level concept (number of documents)	IoT Products (11), Banking (7), Medical Devices (5), CPU (6), Software (5), Hardware (6) Vulnerabilities, Ransomware (75), Healthcare (5)
Equivalent concept (number of documents)	Mirai (13), Dark Web (3), Phishing Emails (41), Targeted Attacks (6), Wannacry (34), Satan (1), JenX (1), Banking Trojan (13), Petya (21), NotPetya (10), Eternal Blue (16), ZDI (16), Bitcoin (4), VMWare (9), Sage (1), APT28 (2), Monero (7), Hajime (2), Satori (1), Wanacrypt0r (3), BadRabbit (4), kirk (1), ATM (5), Cryptolocker (5), Samsam (4), Shadow Broker (6), Jigsaw (1), Lazarus (4), Spectre•Meltdown (5)
(b) Cluster 1 topics subordinate to cluster 10	
Cluster 10_1 (28 items)	
Upper-level concept (number of documents)	Ransomware (53)
Equivalent concept (number of documents)	Wannacry (11), Petya (8), Samsam (7), Phishing (6), Ransomware-as-a-Service (6), Cyberattacks (5), Exploit Kits (4), Attacks on the Healthcare Industry (4), No More Ransom (3), Ryuk (3), ETERNALBLUE (2), Mobile Devices (2), Satan (2), Mikikatz (2), Business Email Breach (1), Sage2.0 (1), Kirk (1), Machine Learning (1), Jaff (1), Crytowall (1), Teslacrypt (1), Mole (1), Locky (1), Defray (1), CRYPSHED (1), BadRabbit (1), Cryptolocker (1), Backup (1), Cyborg (1)

Table 6 Cluster topics for "Triton"

(a) Cluster 0 topics	
Cluster 0 (244 items)	
Upper-level concept (number of documents)	Targeted attack (36), IoT (30), Mobile Threat (20), antivirus software (20)
Equivalent concept (number of documents)	Triton (4), APT29 (5), APT33 (3), APT38 (7), APT41 (2), Microsoft Vulnerability (5), Attack on Financial Institutions (11), C2 Server (6), Trojan (16), spearfishing (30), healthcare (2), ZDI (10), BitPaymer (1), DDoS (5), LoclPOS (1), Shamoon (5), Shamoon2 (3), Petya/NotPetya (5), Rink (1), Backdoor (17), SMB (4), CARBANK (4), FINSPY (3), Wannacry (23), Miuref (2), EternalBlue (15), KRACK (1), Click2Gov (1), BadRabbit (3), Danabot (1), Mirai (10), VPNFilter malware attack (1), Metamorfo (1), JuiceJacking (1), ZEUS (1), SANNY (1), Satori (1), SNS (5)
(b) Cluster 0 topics subordinate to cluster 6	
Cluster 0_6 (7 items)	
Upper-level concept (number of documents)	triton (6), industrial control system ICS (7)
Equivalent concept (number of documents)	Stuxnet (2), Industroyer (2), OS (4), Triconex Safety Instrumented System(SIS) (3), Backdoor (1), Mimikatz (1), TriStation (2), Python (2), Smart Factory (1), Industry 4.0 (1), IoT (2), IIoT (2), DDoS (1), Smart Device (1), TEMP.Veles (1), Wannacry, NotPetya (1)

However, documents only with Triton-specific content were gathered.

Table 7 shows the results of the topical survey on "Shamoon." While various malware and system vulnerabilities existed in cluster 5 without the proposed method, the topics were too broad to be used for document retrieval. By contrast, (b) sub-cluster 5 provided information on the Greenbug and Timberworm cyberattacker groups, which are believed to be involved in the Shamoon attack, and related documents on the cyberattacker groups Sofacy, Gallmaker, Whitefly, Inception Framework. These cyberattacker groups conduct targeted attacks through cyber espionage and phishing attacks. We also obtained information on malware with destructive activities, such as NotPetya related to the Disttrack malware used in the Shamoon attack. We were able to obtain documents related mainly to Shamoon.

Table 8 shows the results of the topical survey on "Mobile Malware." From Cluster 11, we obtained information on overlay attacks and malware against Android devices such as Fakeapp and Lockdroid.E. In addition, there were many documents related to the GDPR, the personal data protection law enacted in the EU. In particular, in 2017, the year after the GDPR was enacted, there were a number of GDPR documents. The majority of these documents were reminders that mobile apps must comply with the GDPR if they are

Table 7 Cluster topics for "Shamoon"

(a) Cluster 5 topics	
Cluster 5	
Upper-level concept (number of documents)	Phishing (149), ransomware (25)
Equivalent concept (number of documents)	Targeted Attacks (17), wannacry (34), Petya (16), Vulnerabilities in Websites (7), WordPress (4), Cobalt Strike (1), Atom Bombing (1), Trojan (7), Sathubot (1), ddos (6), APT28 (5), Wannacry (4), Dridex/Locky (1), Rammit (1), Cryptocurrency (4), User Training (4), Petya/NotPetya (3), HawkEye (1), Office365 Vulnerability (7), botnet (2), Cloud (3), XGen (1), SupplyChainAttack (3), Fileless Malware (3), AI/ML (8), Typosquatting (1), IoT (2), Data Breach (2), Blue Coat (1), DMARC (2), TAA (3), NLP (1), GhostMiner/Bluwimps (1)
(b) Cluster 5 topics subordinate to cluster 5	
Cluster 5_5	
Upper-level concept (number of documents)	Cyberattacker group (13), targeted attack (10)
Equivalent concept (number of documents)	Spear Phishing (6), DDoS (6), IoT/IoT Product Vulnerabilities (6), Credential Capture (6), Botnets (6), ML/AI (5), Supply Chain (5), Disruptive Malware (5), Disttrack Malware (4), Shamoon2 (4), Shamoon (3), Spam (4), Living Off The Land (LotL) (3), Mirai (3), NotPetya (3), Middle East (3), Financial Trojan (3), Mimikatz (2), CCleaner attack (2), RAT (1), Gallmaker (2), Whitefly (1), Inception Framework (1), Sofacy (1), Pawn Storm (1), Greenbug (1), Timberworm (1), Credential Stuffing (1), Trojan.Nancrat (1), Trojan.Filerase (1), Hancitor malware (1), Trojan.Ismdoor (1), Trojan.Bachosens (1)

Table 8 Cluster topics for "Mobile Malware"

Cluster 11	
Upper-level concept (number of documents)	Mobile Malware (36), Ransomware (20), GDPR, Data Privacy Protection (59)
Equivalent concept (number of documents)	Android (25), iOS (5), fake apps (6), third party (5), spam (3), TLD (6), Office365 (9), Android.Fakeapp (4), stalkerware (2), Android.Lockdroid.E (3), DDoS (3), Lockdroid.E (3), DDoS (3), Android 8.0 (O) (5), IoT (2), overlay (5), APEC/CBPR (2), Twitter (5), Facebook (6), Rammit (1), spear phishing (2), APT-C-23 (2), APT28 (1), APT34 (1), cryptocurrency (2), Cryptjacking(2), Reputation.1 (2),

to be used in the EU. This suggests that the GDPR has had a significant impact on mobile products.

5.2 Discussion

If there is a label that adequately describes the topic of a document, it can be used as the next search query for security information. The labels in Tables 2, 3, 4, 5, 6, 7 and 8, which are the outputs of the seven cases in Subsection 5.1, can be used as the next search query for in-depth investigation.

Although this study was conducted on 2386 security reports, when building a search engine using the proposed method, the number of documents added to the database will increase as the operation proceeds. A dataset with numerous documents becomes a set of security reports with a mixture of various topics that last for a short, medium, and long period of time. To deal with this, it becomes increasingly difficult to set an appropriate number of topics, generate document vectors, and assign multi-labels in a batch in terms of computational complexity. Therefore, when dealing with a huge dataset, we can follow the approach of [18] and divide the dataset into time periods and apply the proposed method. It has been shown that the partitioning method in [18] can suppress the dispersion of topics, generate appropriate vectors, and assign multi-labels. We believe that our proposed method will not have any problem in terms of performance

and relevance even if the number of documents increases using the partitioning method.

6 Conclusion

To improve the quality of multi-labeling for security reports, we propose a method for extracting summary documents that contain terms related to multiple topics, re-vectorizing them, and finally re-labeling them. By applying the proposed method to 2386 security reports, we confirmed that the summary documents could be vectorized into more detailed topics and the labels could be assigned relatively close to the contents of the documents.

In the proposed method, clustering is used to locate summary documents, and multi-labels for documents are basically assigned by topic distribution Θ and word distribution Φ obtained through LDA. In the case study discussed in Sect. 5, we assumed the output of a search engine that stores multi-label and document vectors and analyzed the topics by treating the clustering results as search results. We have shown that the proposed method can be applied to multi-label and document vectors to collect keyword-specific documents. Without Steps 6 and 7 in the proposed method, which do not use the proposed method, the topics were broadened to include vulnerabilities of systems and products, attacks on various institutions, and various ransomware related to attacks on IoT products, a high-level concept for Mirai, made the vectorization difficult to use for search. By using the method proposed in this paper, it is possible to retrieve documents with contents limited to the keywords of interest, such as information on IoT via the information on IoT products, a high-level concept for Mirai, and documents related to Gafgyt and its variant Hakai, which are equivalent concepts for Mirai. In the future, we hope to implement a retrieval method that combines labels assigned by multiple methods and document vectors, and evaluate its usability. For a higher level of analysis, we will conduct a usability measurement. Another challenge is to clarify how we can measure the usability of our solution in the absence of an evaluation dataset for security information retrieval.

Acknowledgements This research was conducted under a contract of "Research and development on IoT malware removal/make it non-functional technologies for effective use of the radio spectrum" among "Research and Development for Expansion of Radio Wave Resources (JPJ000254)," which was supported by the Ministry of Internal Affairs and Communications, Japan.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 993–1022
2. Rose S, Engel D, Cramer N, Cowley W (2010) Automatic keyword extraction from individual documents. In: Berry MW, Kogan J (eds) Text mining: Applications and theory. Wiley Online, pp 1–20
3. Mihalcea R, Tarau P (2004) TextRank: bringing order into texts. In: Proc. the 2004 Conference on Empirical Methods in Natural Language Processing, pp 404–411
4. Witten IH, Paynter GW, Frank E, Gutwin C, Nevill-Manning CG (1999) KEA: practical automatic keyphrase extraction. In: Proc. the Fourth ACM Conference on Digital Libraries, pp 254–255
5. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC (2001) Estimating the support of a high-dimensional distribution. Neural Comput 13(7):1443–1471. https://doi.org/10.1162/089976601750264965
6. Arun R, Suresh V, Veni Madhavan CE, Narasimha Murthy MN (2010) On finding the natural number of topics with latent dirichlet allocation: some observations. In: Lecture Notes in Computer Science, pp 391–402. https://doi.org/10.1007/978-3-642-13657-3_43
7. Mimno D, Wallach HM, Talley E, Leenders M, McCallum A (2011) Optimizing semantic coherence in topic models. In: Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11). ACM, pp 262–272
8. tmtoolkit PyPl. https://pypi.org/project/tmtoolkit/. Accessed 2020/12/11
9. Simply security news, views and opinions from Trend Micro https://blog.trendmicro.com/. Accessed 2020/12/11
10. Cisco blog, from https://blogs.cisco.com/. Accessed 2020/12/11
11. Symantec blogs. https://www.symantec.com/blogs/. Accessed 2020/12/11
12. Barracuda- security, access and reliability for cloud-connected networks and applications. https://blog.barracuda.com/. Accessed 2020/12/11
13. Druva blog: Data protection and beyond. https://www.druva.com/category/tech-engineering/. Accessed 2020/12/11
14. Threat research. https://www.fireeye.com/blog/threat-research.html. Accessed 2020/12/11
15. Network security blog. https://www.netscout.com/asert. Accessed 2020/12/11
16. Palo alto networks blog. https://blog.paloaltonetworks.com/. Accessed 2020/12/11
17. GuidedLDA: semi supervised guided topic model with custom guidedLDA. https://github.com/vi3k6i5/. Accessed 2020/12/11
18. Nagasawa R, Furumoto K, Takita M, Shiraishi Y, Takahashi T, Mohri M, Takano Y, Morii M (2021) Partition-then-overlap method for labeling cyber threat intelligence reports by topics over time. IEICE Trans Inf Syst E104.D(5):556–561. https://doi.org/10.1587/transinf.2020DAL0002