セキュリティレポートのマルチラベル分類のための トピックモデルの汎化性能に着目した外れ値検出の適用

長田 侑樹[†1] 瀧田 愼[†2] 古本 啓祐[†3] 白石 善明[†1]
高橋 健志[†3] 毛利 公美[†4] 高野 泰洋[†1] 森井 昌克[†1]

概要：セキュリティレポートに付与されるラベルは発行元によって異なっており，自組織に関連するレポートを見つけ出すのは容易ではない．レポートの内容に応じて一元的にラベル付けをすることができれば，セキュリティオペレーターなどが所望のセキュリティレポートを得るための助けになる．本稿では，トピックモデルによりセキュリティレポートのマルチラベル分類を行うときの文書に付与されるマルチラベルの質の向上を目的として，内容が他と異なる文書を外れ値文書と見なし，外れ値文書を除いてトピックモデルを構築することを提案する．ケーススタディとしてセキュリティベンダー8 社が 2017 年から 2019 年に発行した 2386 件のセキュリティレポートに対して提案手法を適用し，トピックモデルの評価値である Perplexity を考慮することでトピックモデルの汎化性能が向上することを確認した．また，汎化性能の向上に伴い，クラスタリングの結果が良くなることを確認した．

キーワード：脅威情報，マルチラベリング，トピックモデル，外れ値検出，セキュリティレポート

Applying Outlier Detection to Improving Modeling Accuracy of Topic Models for Multi-Labeling Security Reports

Yuki Osada[†1] Makoto Takita [†2] Keisuke Furumoto [†3] Yoshiaki Shiraishi [†1]
Takeshi Takahashi [†3] Masami Mohri [†4] Yasuhiro Takano [†1] Masakatu Morii [†1]

Abstract: Because the labels given to security reports vary from one security vendor to another, finding the reports associated with the incidents is not easy. So, if we can label the reports appropriately according to their content, it will help the security operators to get suitable security reports. In this paper, we propose a method for constructing a topic model for multi-label classification of security reports, excluding outlier documents whose contents are different from other documents in order to improve the quality of multi-labels. As a case study, the proposed method was applied to 2386 security reports published by eight security vendors from 2017 to 2019. The results showed that considering the evaluation value of the topic model (Perplexity) improved the generalization performance and the clustering results.

Keywords: Threat Intelligence, Multi-Labeling, Topic models, Outlier Detection, Security Reports

1. はじめに

インターネットのインフラ化に伴い，組織を標的としたサイバー攻撃が高度化・複雑化している．組織はその環境に応じてインシデント発生前の事前対応を行うとともに，もしもインシデントが発生したらその事後対応を迅速に行わなければならない．事前対策及び事後対応には，インシデントの発生原因や攻撃手法などがまとめられたセキュリティレポートが有用である．しかし，セキュリティベンダー各社が定期的に発行しているセキュリティレポートは日々増加する一方であり，自組織に応じたレポートを適切に探し出すことは容易ではない．セキュリティレポートにラベルを付ける基準は統一されておらず，発行元によって様々である．ラベルが付けられていないレポートもある．つまり，レポートに付与されているラベルを用いて複数の情報源からのレポートの分類は困難である．レポートの内容に応じて一元的にラベル付けをすることができれば，セキュリティオペレーターなどが所望のセキュリティレポートを得るための助けになる．セキュリティレポートには１つの文書内にキーワードとなるような重要単語やトピックが複数出現するためマルチラベルが適している．

トピックモデルはクラスタリングにおける次元削減のアプローチとして注目されている．Blei らにより提案された LDA(Latent Dirichlet Allocation)[1]はトピックモデルの一種で，文書内の単語の共起性によって文書中に現れない潜在的な話題を解析することができる．筆者らはトピックモデルを用いてセキュリティレポートにマルチラベルを付与する手法を提案している[2]．それは，まず，セキュリティレ

†1 神戸大学 大学院工学研究科
†2 兵庫県立大学 社会情報科学部
†3 国立研究開発法人 情報通信研究機構
†4 岐阜大学 工学部電気電子・情報工学科

ポートの文書ベクトルをトピックモデルで次元圧縮し，その低次元ベクトルをクラスタリングすることで潜在的な話題が似ている文書を集め，次に，各クラスタに含まれる文書に対してトピックモデルのパラメータに基づきマルチラベルを付与するという二つの方法を組み合わせた手法である．付与されたマルチラベルの良否は，前者のクラスタリングの結果，すなわちトピックモデルの汎化性能が影響する．もしもモデル構築の入力データの内容がマルチラベルによる検索結果として望ましいものと大きく異なるものが多数存在する場合，トピックモデルの汎化性能が低くなり，結果的にマルチラベルの質が悪くなると考えられる．

本稿では，文書に付与されるマルチラベルの質の向上を目的として，内容が他と異なると考えられる文書を“外れ値文書”と見なし，外れ値文書を除いてトピックモデルを構築することを提案する．セキュリティベンダー8 社が 2017 年から 2019 年に発行した 2386 件のセキュリティレポートに対して提案手法を適用し，トピックモデルの評価値である Perplexity を考慮することでトピックモデルの汎化性能が向上することを確認している．そして，汎化性能の向上に伴ってクラスタリングの結果がよくなることを確認している．

2. トピックモデルと外れ値検出

2.1 トピックモデル

トピックモデルは潜在意味解析を行う手法の一つで，文書内の単語の共起性により潜在的な話題を解析することができる．トピックモデルの一種である LDA は，各文書に潜在的に複数のトピックが存在すると仮定し，文書を単語の集合と捉え，単語の共起情報からトピックを推定することで文書中に現れない話題を解析する．LDA に文書ベクトルを入力しトピック数を指定することでトピックの生成が行われ，文書は各トピックへの所属確率 θ でベクトル化される．また，各トピックはトピックを構成する単語集合の確率分布 ϕ でベクトル化される．

2.2 トピックモデルの評価指標

トピックモデルの評価指標として Perplexity[3] と Coherence[4]が用いられる．学習データでトピックモデルの学習を行い，未知データに対するモデルの予測精度を測る．Perplexity はトピックモデルの汎化性能を測る評価指標である．Perplexity 値が低いほど，高い精度で予測できる汎化性能の高いトピックモデルを意味する．Coherence はトピックを構成する単語集合の意味的な一貫性を測る評価指標である．Coherence 値が高いほど，トピックを構成する単語集合の意味が一貫しており人間にとって解釈しやすいトピックであることを意味する．

2.3 外れ値検出

本稿で用いる教師なしの外れ値検出手法を説明する．以降では外れ値の対義語を“通常値”と呼んでいる．

2.3.1 Isolation Forest

Isolation Forest (iForest) [5]は，データセットを二分探索木に編成する際，外れ値は通常値に比べてデータ数が少ないため比較的浅い段階で分割される確率が高くなるという考え方による外れ値検出アルゴリズムである．まずデータセットからランダムにサンプルを n 個取得する．取得した各サンプルに対してランダムに分割値を選択することで二分探索木を構築する．それぞれのサンプルの外れ値スコアはすべての木の平均深さから導出され，外れ値は通常値と比べて平均深さが短く，スコアが高くなる．

2.3.2 Local Outlier Factor

Local Outlier Factor (LOF)[6]は局所密度の概念に基づいており，あるオブジェクトの局所密度をその近傍群の局所密度と比較することで，密度が同程度である点と密度が低い点を特定することができる．前者を通常値，後者を外れ値として分類する．

2.3.3 One-Class SVM

One-Class SVM[7] は Support Vector Machine(SVM)の拡張アルゴリズムであり，1 クラス分類を行う手法である．データセットを非線形カーネル関数によって入力空間から特徴空間へマッピングし，通常値と外れ値に分離する識別境界を引くことで外れ値検出を行う．

3. 提案手法

モデルの汎化性能を向上させるために，内容が他と大きく異なるセキュリティレポートを除いてトピックモデルを構築する方法は以下のとおりである．

Step 1. 前処理

セキュリティレポートの前処理として，各レポートからタイトル，本文，図表のキャプションを抽出し，単語に分割する．また，セキュリティレポートには「browser base」や「mobile device」などのセキュリティ業界の専門用語が現れるので，文書を連続した N 個の単語で分割するテキスト分割方法である N-gram の N=2 を用いて複合語とすることでセキュリティレポート中の専門用語抽出を行う．また，前置詞や冠詞，代名詞などのストップワードに加え，マルウェアのハッシュ値や URL，C&C サーバの IP アドレスなど本文とは関係のない記述を除去する．単語の出現頻度を基に文書をベクトル化する BoW(Bag-of-Words) を用いて，単語・複合語の出現回数を基に各文書をベクトル化する．

Step 2. 外れ値検出

各アルゴリズムのパラメータは初期設定とし，文書ベクトルに外れ値検出アルゴリズムを適用する．外れ値とされた文書を“外れ値文書”，それ以外の文書を“通常値文書”とする．

表 1 トピックモデルの汎化性能評価

	Model (外れ値検出なし)	Model (Isolation Forest)	Model (LOF)	Model (OneClassSVM)
1	トピック数:32 Coherence:-1.037 Perplexity:2437.3	トピック数:20 Coherence:-0.582 Perplexity:**2101.8**	トピック数:32 Coherence:-1.278 Perplexity:2640.9	トピック数:14 Coherence:-2.008 Perplexity:**2363.7**
2	トピック数:42 Coherence:-1.419 Perplexity:2852.7	トピック数:28 Coherence:-0.975 Perplexity:**2469.3**	トピック数:28 Coherence:-1.017 Perplexity:**2624.1**	トピック数:14 Coherence:-0.971 Perplexity:**2372.7**
3	トピック数:34 Coherence:-0.838 Perplexity:2508.4	トピック数:28 Coherence:-0.912 Perplexity:**2375.5**	トピック数:34 Coherence:-1.146 Perplexity:2717.0	トピック数:16 Coherence:-2.361 Perplexity:2545.2
4	トピック数:34 Coherence:-0.838 Perplexity:2554.7	トピック数:24 Coherence:-0.661 Perplexity:**2401.5**	トピック数:30 Coherence:-1.084 Perplexity:2662.2	トピック数:14 Coherence:-2.361 Perplexity:**2413.5**
5	トピック数:32 Coherence:-0.915 Perplexity:2607.8	トピック数:28 Coherence:-0.906 Perplexity:**2603.1**	トピック数:28 Coherence:-1.039 Perplexity:2703.1	トピック数:12 Coherence:-0.715 Perplexity:**2254.0**

Step 3. 最適なトピック数の決定

本稿ではトピックモデルとして LDA を用いる．LDA でトピックモデルを構築する際，ハイパーパラメータであるトピック数を事前に指定する必要がある．本稿では，Python の tmtoolkit パッケージ[8]を用いてトピック数を自動決定する．本パッケージで使用可能なトピック数の評価手法である Arun_2010[9], Coherence_mimno_2011[4]を用いてトピック数を決定する．はじめに，いくつかのトピック数の候補を入力して各トピック数における評価値を算出する. 次に, Arun_2010 は最小値を最適とし，Coherence_mimno_2011 は最大値を最適とする評価指標であるため, Coherence_mimno_2011 の値を正負反転することで共に最小値を最適とする処理を行う．そして最大値を 1，最小値を-1 とする正規化を行い，各トピック数における評価値の平均値を求める．平均値が最小となるトピック数を最適なトピック数とする．

Step 4. トピックモデルの学習と文書のベクトル化

トピック数と通常値文書を LDA に入力し，トピックモデルの学習を行う．その後，外れ値文書を含めた全文書をトピックモデルに入力し，各文書をトピック分布 θ でベクトル化する．また，同時に各トピックの単語分布 ϕ を得る．

以上の手順の後に，文献[2]の手法にもとづいて，文書ベクトル θ を用いてクラスタリングを行い，そして単語分布 ϕ を用いたマルチラベルを付与する．

4. 汎化性能の評価

4.1 データセットと実験環境

セキュリティベンダー8 社(Trendmicro[10], Cisco[11], Symantec[12], Barracuda[13], Druva[14], FireEye[15], Arbor[16], Paroalto[17])が 2017 年 1 月 1 日～2019 年 12 月 31 日に発行したセキュリティレポート 2386 件をデータセットとした．LDA の構築には Python3 の GuidedLDA ライブラリ[18]を用いた．ディリクレ分布のパラメータ α, β の値は共に 0.01 とし，クラスタ数はトピック数と同じ数を設定した．クラスタリングは k-means を用いた．外れ値検出アルゴリズム及び Perplexity 値算出は scikit-learn ライブラリを利用した．

4.2 トピックモデルの汎化性能評価

提案手法の有用性を確認するため，外れ値検出を適用したトピックモデルと適用しないトピックモデルを 3 種類構築し，Perplexity による評価を行った．まずデータセットを文書数が等しくなるようにランダムに 5 分割し，4/5 を訓練データとしトピックモデルを構築する．その際，外れ値検出を適用せず全訓練データで学習を行うモデルと，外れ値検出を適用して通常値文書だけで学習を行うモデルを作成する．そして 1/5 の評価データをそれぞれのモデルに入力して Perplexity 値を算出する．評価データを入れ替えて交差検証を行った．その結果を表 1 に示す．iForest, One-Class SVM において，外れ値検出を適用しなかったモデルに比べて外れ値検出を適用したモデルにおける Perplexity 値が低くなっている．これは外れ値文書を省いてトピックモデルを構築することで単語の平均分岐数が少ないモデルを構築できるということである．提案手法によって汎化性

表 2　参照文書が最も集中しているクラスタ
（外れ値検出は One-Class SVM）

Wannacry	クラスタ番号(クラスタ内の参照文書数/全参照文書数)
外れ値検出なし	クラスタ 22 (48 件/159 件)
外れ値検出あり	クラスタ 10 (**61 件/159 件**)

Shamoon	クラスタ番号(クラスタ内の参照文書数/全参照文書数)
外れ値検出なし	クラスタ 14 (9 件/18 件)
外れ値検出あり	クラスタ 5 (**10 件/18 件**)

Mirai	クラスタ番号(クラスタ内の参照文書数/全参照文書数)
外れ値検出なし	クラスタ 5 (32 件/57 件)
外れ値検出あり	クラスタ 10 (30 件/57 件)

Triton	クラスタ番号(クラスタ内の参照文書数/全参照文書数)
外れ値検出なし	クラスタ 4 (4 件/9 件)
外れ値検出あり	クラスタ 0 (**7 件/9 件**)

Samsam	クラスタ番号(クラスタ内の参照文書数/全参照文書数)
外れ値検出なし	クラスタ 7 (11 件/19 件)
外れ値検出あり	クラスタ 10 (**13 件/19 件**)

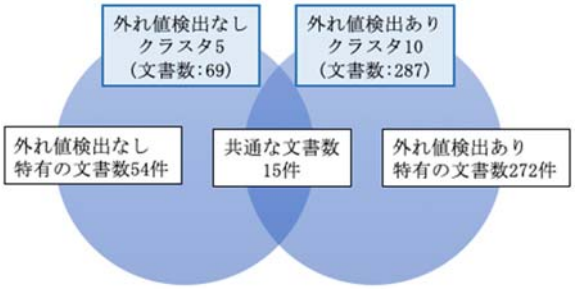
能の高いモデルが構築可能であることが確認された.

4.3 参照文書によるクラスタリングの評価

汎化性能向上によるクラスタリングの結果の変化を見る．ここでは，本文中に特徴的な単語を含む文書を“参照文書”と呼ぶこととする.参照文書は内容が類似しており，文書クラスタリングによって同じクラスタに分類されることがと期待されるため，参照文書の密集度合いを測定することでクラスタリングの良否を測る.

まず全文書中で「Wannacry」,「Shamoon」,「Mirai」,「Triton」,「Samsam」というキーワードを含む文書をそれぞれの参照文書とした．提案手法によって汎化能力が最も向上した表 1 のテスト 5 のデータに One-Class SVM を適用したモデルを用いた．そして 4.2.節と同様に，全ての訓練データで学習を行うモデルと One-Class SVM を適用し通常値文書だけで学習を行うモデルそれぞれでクラスタリングを行う.

それぞれの参照文書が最も集中しているクラスタにおける参照文書の数を表 2 に示す.「Wannacry」というキーワードに着目した際，外れ値検出適用前後でクラスタ 22 とクラスタ 10 は「Wannacry」の参照文書が最も集中しており，意味的に類似したクラスタだと考えられる.「Wannacry」



(a)　包含関係に基づく分類：「Mirai」

「Mirai」	キーワード含む	キーワード含まない
共通な文書数（15 件）	15	0
クラスタ 5 特有の文書数（54 件）	17	37
クラスタ 10 特有の文書数（272 件）	13	259

(b)　キーワード「Mirai」に基づく分類

図 1　「Mirai」に関するクラスタの内訳

が含まれる文書は全文書中 159 件存在し，外れ値検出なしのモデルでは 48 件がクラスタ 22，外れ値検出ありのモデルでは 61 件がクラスタ 10 に分類された．外れ値検出したモデルの方が参照文書が一つのクラスタに集中しており，クラスタリング精度の向上が確認できた.

5. ケーススタディ：外れ値検出適用でクラスタが変化した文書の分析

表 3 より「Mirai」というキーワードに着目した際，外れ値検出適用前のクラスタ 5 と適用後のクラスタ 10 は「Mirai」に関係する文書の集合と考えた．外れ値検出適用前後のクラスタ内の文書の変化を調べるために，それぞれのクラスタ内の文書が扱う話題を調査する.

それぞれのクラスタ内の文書を包含関係をもとに分類した上で，キーワード「Mirai」が存在する文書と存在しない文書に分類した．その結果を図 2 に示す．それぞれのクラスタに共通する文書は外れ値検出適用による影響を受けていない.また,キーワード「Mirai」が含まれる文書は「Mirai」に関する話題を持つとわかる．そこで，それぞれのクラスタに特有かつ，キーワードが含まれない文書を対象として話題を調査する.

それらの文書は，キーワードが含まれず「Mirai」と別の話題を持つが，クラスタリングによって「Mirai」と同じクラスタに分類された文書である．それらを分析することで「Mirai」と意味的に近い文書を得ることができると期待される．分析対象文書はクラスタ 5 の 37 件，クラスタ 10 の 259 件である．それぞれの文書の話題を目視で確認し，キ

表 3　「Mirai」のクラスタの話題

(a)　外れ値検出なしのクラスタ 5 の話題

クラスタ 5	
上位概念 (文書数)	IoT 製品の脆弱性(29)
等位概念 (文書数)	ランサムウェア(8), ヘルスケア(6), Operating Technology(4), VPNFilter malware attack(4) , BruneBorne(1), SSDP Diffraction attack(3)

(b)　外れ値検出ありのクラスタ 10 の話題

クラスタ 10	
上位概念 (文書数)	ランサムウェア(75), ヘルスケア(5), IoT 製品(11)・銀行(7)・医療機器(5)・CPU(6)・ソフト(5)・ハード(6)の脆弱性
等位概念 (文書数)	ダークウェブ(3), フィッシングメール(41), 標的型攻撃(6), wannacry(34), Satan(1), Banking Trojan(13), Petya(21), Eternal Blue(16), ZDI(16), Bitcoin(4), VMWare(9), Sage(1), Monero(7), Wanacrypt0r(3), BadRabbit(4), ATM(5), Cryptolocker(5), Samsam(4), Shadow Broker(6), APT28(2), Jigsaw(1), Lazarus(4), kirk(1), Spectre・Meltdown(5), NotPetya(10)

ーワード「Mirai」に関して上位概念と等位概念に分類した結果を表 3 に示す.

クラスタ 5, クラスタ 10 には共に上位概念として IoT の脆弱性に関する文書が多く存在する. 両クラスタは「Mirai」に関する文書集合であるため妥当な結果である．また，両クラスタ共に, ヘルスケアに関する文書が存在した.「Mirai」と同じクラスタに分類された原因として，クラスタの上位概念である IoT 製品とヘルスケア業界で用いられる医療用機器が関連することやヘルスケア業界を標的として Mirai が使われていることなどが推測できる.

クラスタ 10 では Mirai の等位概念として，Wannacry, Petya などのより広義にマルウェアに関する文書が集まっている．上位概念であるランサムウェアから派生して Wannacry や NotPetya などもたどれることになる. そして，それぞれのマルウェアに紐づいて，IoT 以外にも上位概念が広がっていることがわかる.

表 4 は「Triton」というキーワードに着目した際，外れ値検出適用前のクラスタ 4 と適用後のクラスタ 0 は「Triton」に関係する文書の集合と考えたものである．表 5 はそれぞれの文書の話題を目視で確認し，キーワード「Triton」に関して上位概念と等位概念に分類した結果である．外れ値検出なしのクラスタ 4 は IoT でまとまっているが，外れ値検出有りのクラスタ 0 では標的型攻撃が上位概念として含まれ, それにつながって APT38 などの攻撃者集団の情報をたどれることになる.

このことから，適切なマルチラベルを付与することで，キーワードによる検索に留まらない高度な検索が可能になると期待できる.

表 4　キーワード「Triton」に基づく分類

「Triton」	キーワード含む	キーワード含まない
共通な文書数（20 件）	3	17
クラスタ 4 特有の文書数（194 件）	1	193
クラスタ 0 特有の文書数（244 件）	4	240

6. おわりに

本稿では，セキュリティレポートに付与するマルチラベルの質の向上を目的として，汎化性能を上げるために外れ値検出を適用してトピックモデルを構築する手法を提案した．セキュリティレポート 2386 件を対象に提案手法を適用したところ，汎化性能の高いトピックモデルを構築できることが確認された．そして，マルチラベルを付与する単位となる文書クラスタが汎化性能の向上に伴ってどのように変化するか確認した．具体的には，特徴的なキーワードを持つ参照文書の密集度合いを比較することで，提案手法を適用した方が良好なクラスタを形成できていることを確認した.

提案手法を適用することで注目する特徴的なキーワードが含まれない文書でも内容の似ている文書としてより多く集められることになった．例えば，Mirai の上位概念を IoT 製品，IoT 製品を介して Mirai の等位概念となる EternalBlue が関連する文書として見られるなど，上位概念を介して注目するキーワードを含まない等位概念の内容の文書を取り出すことが可能となることが確認できた.

提案手法を適用しない場合は，特徴的な話題を持つ文書が集まる結果になっている．この結果は文書検索のための一般的に期待される結果であり，特定の話題に絞って検索をしたい場合には外れ値検出をしないモデルにより付与したラベルを使えばよい．それに対して，特定の話題から広がりを持った検索をしたい場合には外れ値検出を適用したモデルにより付与したラベルを用いることが有用であると示唆する結果が得られた．これらの組み合わせにより単語検索に留まらない検索が可能になると期待される．異なる方法で付与したラベルを組み合わせた高度な検索手法の実現と評価を今後行いたい.

表 5 「Triton」に関するクラスタの話題

(a) 外れ値検出なしのクラスタ 4 の話題

クラスタ 4	
上位概念(文書数)	サイバーセキュリティ(40), クラウドの保護(19), インフラセキュリティ(13), IoT(37), Mobile Security(21), OT(10)
等位概念(文書数)	電子メールのセキュリティ(15), ランサムウェア(7), ゼロデイ攻撃(6), ヘルスケア(13), フィッシング(5), SQL(2), Car Hacking(3), SaaS(3), AI(17), iOS/Android の脆弱性(4), Shodan(2), BEC(7), Pll(4), DDoS(4), Wannacry(10), Xgen(3), Mirai(3), Petya(2), AR/VR(2), DX(3), SunOrcal(1), Blaster(1), Supply Chain Attack(3), MIMIKATZ(1), Dragonfly(1), SNS(3)

(b) 外れ値検出ありのクラスタ 0 の話題

クラスタ 0	
上位概念(文書数)	Mobile Threat(20), ウイルス対策ソフト(20), 標的型攻撃(36)
等位概念(文書数)	マイクロソフト脆弱性(5), 金融機関への攻撃(11), C2 サーバー(6), トロイの木馬(16), スピアフィッシング(30), ヘルスケア(2), ZDI(10), IoT 製品(30), BitPaymer(1), APT29(5), APT33(3), APT38(7), APT41(2), DDoS(5), LoclPOS(1), Shamoon(5), Shamoon2(3), Petya/NotPetya(5), Ryunk(1), Backdoor(17), SMB(4), CARBANK(4), FINSPY(3), Wannacry(23), Miuref(2), EternalBlue(15), KRACK(1), UDP Reflection/Amplification DDoS Attacks (1), SSDP reflection/amplification attacks(1), Click2Gov(1) , Danabot(1), VPNFilter malware attack(1), Metamorfo(1), JuiceJacking(1), ZEUS(1), SANNY(1), Satori(1), Mirai(10), BadRabbit(3), SNS(5)

謝辞 本研究は国立研究開発法人情報通信研究機構の委託研究「機械学習に基づくサイバー攻撃情報分析基盤技術の研究開発」により行われた.

参考文献

[1] D.M. Blei, A.Y. Ng and M.I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, p.993-1022.

[2] 長田侑樹，瀧田愼，古本啓祐，白石善明，高橋健志，毛利公美，高野泰洋，森井昌克，“トピックモデルとクラスタリングによるセキュリティレポートのマルチラベル分類”，電子情報通信学会技術研究報告，vol.119，no.437，ICSS2019-100，pp.283-288，2020 年.

[3] H.M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. ICML, 2009b

[4] D. Mimno, H. Wallach, E. Talley, M. Leenders, A. McCullum. Optimizing semantic coherence in topic models. EMNLP, 2011.

[5] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining, 2008, p. 413-422.

[6] M.M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander. Lof: Identifying density-based local outliers. SIGMOD Conference, 2000, p.93-104.

[7] B. Scholkopf, J. Platt, J. Shawe-Taylor, A.J. Smola, and RC Williamson. Estimating the support of a high-dimensional distribution. Neural Computation, 2001, p.1443-1471.

[8] “tmtoolkit・PyPl”. https://pypi.org/project/tmtoolkit/, (参照 2020-08-16).

[9] R. Arun, V. Suresh, C. E. Veni Madhavan, M. N. Narasimha Murthy. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. PAKDD 2010: Advances in Knowledge Discovery and Data Mining, 2010, pp 391-402.

[10] “Simply Security News, Views and Opinions from Trend Micro”. https://blog.trendmicro.com/, (参照 2020-08-16).

[11] “Cisco Blog”. https://blogs.cisco.com/. (参照 2020-08-16).

[12] “Symantec Blogs”. https://www.symantec.com/blogs/. (参照 2020-08-16).

[13] “Barracuda- Security, Access and Reliability for Cloud-Connected Networks and Applications”. https://blog.barracuda.com/. (参照 2020-08-16).

[14] “Druva Blog: Data Protection and Beyond”. https://www.druva.com/category/tech-engineering/. (参照 2020-08-16).

[15] “Threat Research”. https://www.fireeye.com/blog/threat-research.html. (参照 2020-08-16).

[16] “Network Security Blog”. https://www.netscout.com/asert. (参照 2020-08-16).

[17] “Palo Alto Networks Blog”. https://blog.paloaltonetworks.com/. (参照 2020-08-16).

[18] “GuidedLDA: semi supervised guided topic model with custom guidedLDA”. https://github.com/vi3k6i5/. (参照 2020-08-16).