トピックモデルを用いたセキュリティレポートのマルチラベリングのための分割重複入力

長澤 龍成[1] 古本 啓祐[2] 瀧田 愼[3] 白石 善明[1]
高橋 健志[2] 毛利 公美[4] 髙野 泰洋[1] 森井 昌克[1]

概要：日々増加していくセキュリティレポートから，セキュリティオペレーターが所望の情報を得やすくするためには，内容に応じて適切なマルチラベルを付与することが望まれる．大規模かつ不均質な文書集合から何かしらの情報を獲得するための統計的モデリング手法の一つとしてトピックモデルがあり，マルチラベリングに用いることができる．トピックモデルではトピック数が主要なハイパーパラメータであり，単調増加するセキュリティレポートに対してトピック数を適切に変更していく単純な方法では，ハイパーパラメータの探索と再学習の時間が増えるばかりである．適切なラベリングには一般に入力数が大きくなるにしたがってトピック数を大きくすることになる．しかし，トピック数を大きくしても，ラベルとしてふさわしいある時期に局所的に出現する用語が捉えにくくなる．そこで本論文では，入力数が増えてもトピック数を大きくしないという方針の下に，入力データセットを重なり期間をもたせて一定期間で分割してトピックモデルに入力する手法を提案する．トピックモデルとして Latent Dirichlet Allocation（LDA）と Topics Over Time（TOT）に提案手法を適用したデータセットを入力した結果，セキュリティレポートのマルチラベリングとしてふさわしい単語を含むトピックを多く抽出できることが示唆された．

キーワード：トピックモデル， Latent Dirichlet Allocation，Topics Over Time，マルチラベリング，セキュリティレポート

Partition-Then-Overlap Method for Labeling Cyber Threat Intelligence Reports by Topic Model

Ryusei Nagasawa[1] Keisuke Frumoto[2] Makoto Takita[3]
Yoshiaki Shiraishi[1] Takeshi Takahashi[2] Masami Mohri[4]
Yasuhiro Takano[1] Masakatu Morii[1]

Abstract: In order to make it easier for security operators to obtain the information they need from the increasing number of security reports, it is desirable to assign appropriate multi-labels depending on the content. One of the statistical modeling techniques for assigning multi-labels to a large and heterogeneous set of documents is the topic model. However, the approach of changing the number of topics for an increasing number of security reports using the topic model only increases the time required to compute the hyperparameters. In this paper, we propose a method to divide an input data set with an overlapping period of time, based on the idea that the number of topics does not increase as the number of inputs increases. As a result of inputting a dataset of Latent Dirichlet Allocation (LDA) and Topics Over Time (TOT) as topic models, it is suggested that the proposed method can extract a lot of topics that contain words suitable for multi-labeling of security reports.

Keywords: Topic Model, Latent Dirichlet Allocation, Topics Over Time, Multi Labeling, Security Reports

1. はじめに

セキュリティベンダーが発行するセキュリティレポートには，脅威情報の分析結果や注意喚起が記されている．サイバー攻撃の事前および事後対応のために，その対応方法が示唆されているセキュリティレポートは有用である．しかし，セキュリティレポートとして新たに公表される内容は時々刻々と変化し，その数は増え続ける．その状況下で所望するセキュリティレポートを適切に探し出すことは容易ではない．検索を助けるために，セキュリティレポートにはラベルが付与されていることがあるが，ラベルをつ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．またラベルが付与されていないものも多数ある．広範囲のセキュリティレポートを一元的に検索するための手段は提供されていない．日々増加していくセキュリティレポートからセキュリティオペレーターが所望の情報を得やすくしたいならば，内容に応じて適切なマルチラベルを付与することが必要である．

1 神戸大学
Kobe University
2 情報通信研究機構
National Institute of Information and Communications Technology
3 兵庫県立大学
University of Hyogo
4 岐阜大学
Gifu University

大規模かつ不均質な文書集合から何かしらの情報を獲得するための統計的モデリング手法の一つとしてトピックモデルが提案されている．代表的なものに Latent Dirichlet Allocation（LDA）[1]がある．文書を構成しているトピックを適切に捉えることができれば，マルチラベルを付与できる．しかし，セキュリティレポートを含め，多くの文書集合のトピックの発生は，時間とともに変化する．一般のトピックモデルはトピックの推定に時間を考慮しないため，不明瞭で最適でないトピックが発生する可能性がある．そこで，変動するトピックを捉えることを目的として，時間を明示的にモデル化したトピックモデルである Topics Over Time（TOT）[2]が提案されている．これにより，各文書内の単語を時系列に対応付け，より明示的なトピックを生成することが期待される．

本研究の目的はセキュリティレポートに適切なラベルを付与することである．それに到達するためのアプローチは次の通りである．まず，セキュリティレポートをトピックモデルに入力しトピックを生成する．そして，生成されたトピックから所属確率の高い単語を抽出し，セキュリティレポートに付与するラベルとしてふさわしい単語を調べる．

トピックモデルでは，出現頻度の高い単語を優先的にトピックの構成に用いる．セキュリティレポートの集合から構築したデータセットを単純にトピックモデルに入力すると，ラベルとしてふさわしい固有表現がトピックを構成する単語に含まれる可能性が低くなる場合がある．なぜなら，セキュリティレポートに対するマルチラベリングでは，マルウェア名や攻撃キャンペーン名がラベルに含まれることが期待されるが，そのような固有表現は文書中に多くは含まれない場合があるからである．例えば，攻撃手段や攻撃手法に言及されていても，未だマルウェア名や攻撃キャンペーン名が付けられていないときにこのようなことが起こり得る．したがって，出現回数の少ないセキュリティドメイン固有の表現を抽出することが目的に到達するための解決すべき課題である．なお，ラベルとなる固有表現が含まれていない文書でも，同一のラベルが付与されたレポートは何かしら類似している内容を含んでいる．トピックモデルによるマルチラベリングは単語検索に留まらない情報の獲得，すなわちインテリジェンス情報の抽出が期待できる．

セキュリティレポートを NLTK [3] などの一般的な自然言語処理のツールで前処理をして，頻出単語やストップワードを省いた後も，セキュリティドメインでの常用的なフレーズが残る．入力データが増えるとこの常用的なフレーズの出現頻度が相対的に上昇する．常用的なフレーズは文書集合全体に広く分布している一方で，マルウェア名や攻撃キャンペーンなどの固有表現は局所的に分布している．つまり，文書集合全体に広く分布している単語ではなく，局所的に出現する単語を捉えることが求められる．

また，データセットに追加するセキュリティレポートは年々増加する．データセットの文書数が増えていくにも関わらず，ハイパーパラメータであるトピック数を固定のまま学習すると，特徴的な単語（例えば 4 章の実験結果に少数現れる triton，samsam など）が，例えば上位の概念（マルウェアやランサムウェアなど）にまとまっていく．トピック数を増やすことで特徴的な用語を持つ文書がまとまらないようにできるが，その最適なトピック数を求めることは容易ではなく，またトピック数を求める計算時間は大きくなる．

そこで，本論文では文書数が増えてもトピック数を大きくしないという方針の下に，入力データセットを一定期間で分割して重なり合わせて，トピックモデルに入力する手法を提案する．この手法により，局所的に出現する特徴的な表現が常用単語により埋もれてしまうことや，特徴的な単語が期待しないトピックにまとまることを抑制し，データセットが大きくなっても適切なラベル付けできることが期待される．

2. トピックモデル

2.1 Latent Dirichlet Allocation（LDA）

LDA は各文書の潜在的なトピックから単語が生成されると仮定して，そのトピックを文書集合から推定することを目的としたトピックモデルである．文書中の各単語は独立して存在しているのではなく，それぞれが潜在的なトピックを持ち，同じトピックを持つ単語は同じ文書中に現れやすいという仮定に基づいている．LDA の文書生成過程を表現したグラフィカルモデルを図 1 に示す．LDA は，各文書の単語集合を入力し，図 1 中のθおよびϕを推定する．θはある文書にあるトピックが現れる確率を表し，ϕはあるトピックにある単語が現れる確率を表す．また，α，βはハイパーパラメータ，wは単語，zはトピックを表す．

2.2 Topics Over Time（TOT）

TOT は LDA をベースに考えられたトピックモデルである．LDA ではトピックを推定する際に各文書における単語の共起情報を考慮している．TOT では，それに加えて文書の時間情報を考慮する．文書のトピックを時系列に対応付けることで，共起パターンの混乱や不明瞭で最適でないトピックの発生を抑制できる．TOT の文書生成過程を表すグラフィカルモデルを図 2 に示す．LDA のグラフフィカルモデルに，時間情報を表すtが加えられる．TOT は，各文書の単語集合と文書の時間集合を入力し，θおよびϕに加えて，トピックが時間とともにどのように遷移するかを表すψを推定する．

3. 提案手法：Partition-Then-Overlap Method

解析対象のM個の文書集合において，文書の発行日を順に並べた順序付き時間集合$T = \{t_1, t_2, \dots, t_M\}$と文書を発行日順に並べた順序付き文書集合$D = \{d_1, d_2, \dots, d_M\}$をデー

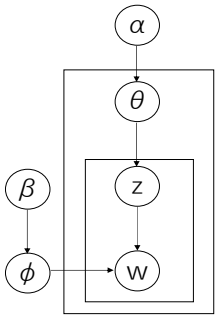

図 1　LDA のグラフィカルモデル

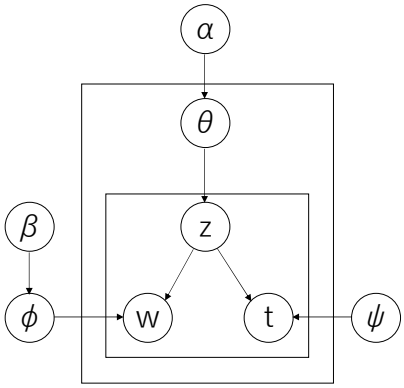

図 2　TOT のグラフィカルモデル

タセットとする．このとき，写像$f: T \to D$は順序を保つ全

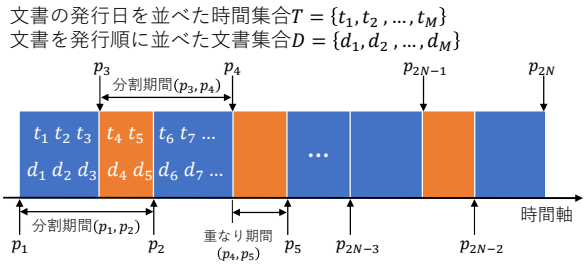

図 3　提案手法によるデータセットの分割

単射であり，TとDは一対一対応する．提案手法では，図 3 のように，一定の分割期間，重なり期間を指定し，データセットである順序集合 T，D をN個に分割する．

$i = 1,2, \ldots, N$に対して，分割期間の両端の日付をp_{2i-1}, p_{2i}とし，分割期間 $(p_{2i-1},\ p_{2i})$ に存在する文書の発行日の集合を $T_{P_i} = \{t_x \in T \mid p_{2i-1} < t_x \leq p_{2i}\}$ とする．T_{P_i} は T の部分集合であり，順序付き集合である．分割期間の集合を $P = \{(p_1, p_2),\ (p_3, p_4),\ \ldots,\ (2_{2N-1}, p_{2N})\}$とする．各分割期間において，$t_x \in T_{P_i}$ に一対一対応する文書$d_x \in D$ の集合を D_{P_i} とする．

$p_{2i+1} < p_{2i}$のとき，分割期間 $(p_{2i-1},\ p_{2i})$ と分割期間 $(p_{2i+1},\ p_{2i+2})$は期間(p_{2i}, p_{2i+1})で日付が重複する．すなわち，分割したデータセットは一定の重なり期間を持つ．このとき，重なり期間の集合は $O = \{(p_2, p_3),\ (p_4, p_5),\ \ldots,\ (2_{2N-2}, p_{2N-1})\}$となる．重なり期間 $(o_{2i},\ o_{2i+1})$ に存在する文書の日付の集合を $T_{O_i} = \{t_x \mid p_{2i} < t_x \leq p_{2i+1}\}$ とする．各重なり期間において$t_x \in T_{O_i}$ と一対一対応する文書$d_x \in D$ の集合を D_{O_i} とする．

分割期間と重なり期間における時間集合の関係は

$$T_{O_i} = T_{P_i} \cap T_{P_{i+1}}$$

となる．また，これより T_{O_i} は以下の大小関係を持つ．

$$0 < \mathrm{Len}(T_{O_i}) < \min\left(\mathrm{Len}(T_{P_i}), \mathrm{Len}(T_{P_{i+1}})\right)$$

Len()は集合の要素数である．文書集合D_{P_i}, D_{O_i} に関しても同様の関係が成り立つ．

分割された時間集合T_{P_i}と文書集合D_{P_i}の順序付き集合族をそれぞれ $T_P = \{T_{P_1}, T_{P_2}, \ldots, T_{P_N}\}$, $D_P = \{D_{p_1}, D_{P_2}, \ldots, D_{P_N}\}$ と表す．順序付き集合族 D_P の各要素に対して以下の前処理を行う．

- 複合語の生成
- ストップワードの除去
- 出現文書率が 50%より高い単語を除去
- 数字，記号，引用符が含まれる単語を除去
- 製品名の除去
- 出現文書数が 3 以下の単語を除去

ある文書d_jにおいて，前処理後に残った単語および複合語により構成された単語集合を $W_j = \{w_1, w_2, \ldots, w_n\}$ とする．分割された文書集合D_{P_i}に含まれる各文書を前処理した単語集合の集合を分割単語集合 $\Delta_{P_i} = (W_1, W_2, \ldots, W_l)$ とする．このとき，分割単語集合 Δ_{P_i} の順序付き集合族は$\Delta_P = (\Delta_{P_1}, \Delta_{P_2}, \ldots, \Delta_{P_N})$ となる．

トピックモデルとして LDA を用いる場合，各文書の単語集合が入力に用いられる．すなわち，LDA の入力は分割単語集合Δ_{P_i}である．ハイパーパラメータα，β，トピック数γ，イテレーション数δを指定し，LDA に集合族Δ_{P_i}を入力する．Δ_{P_i}に対して，文書固有のトピック多項分布θ_i，トピック固有の単語多項分布ϕ_iが出力される．これを集合族Δ_Pの各要素Δ_{P_i}に対して繰り返して，出力集合族$(\{\theta_1, \phi_1\}, \{\theta_2, \phi_2\},\ \ldots,\ \{\theta_N, \phi_N\})$ を得る．

トピックモデルとして TOT を用いる場合，分割単語集合族 Δ_P と分割された時間集合族 T_P を合わせて，入力集合族 $(\{\Delta_{P_1}, T_{P_1}\},\ \{\Delta_{P_2}, T_{P_2}\},\ \ldots,\ \{\Delta_{P_N}, T_{P_N}\})$ を構成する．LDA と同様の初期条件を指定し，入力集合族の各要素を TOT に入力する．　入力集合$\{\Delta_{P_i},\ T_{P_i}\}$ に対して，文書固有のトピック多項分布θ_i，トピック固有の単語多項分布ϕ_i，トピック固有の時間分布ψ_iが出力される．これを入力集合族の各要素に対して繰り返して，出力集合族$(\{\theta_1, \phi_1, \psi_1\},\ \{\theta_2, \phi_2, \psi_2\},\ \ldots,\ \{\theta_N, \phi_N, \psi_N\})$ を得る．

4.　実験と結果

解析対象の文書集合は，セキュリティベンダー8 社（Arbor, Barracuda, Cisco, Druva, FireEye, Paloalto, Symantec,

TrendMicro）のブログページから収集した，2017 年 1 月 1 日から 2019 年 12 月 31 日までのセキュリティレポート 2386 件とする．LDA は教師なしトピックモデリングと自然言語処理のためのオープンソースライブラリである gensim[4] を用いて実装した．TOT は Python3 および拡張モジュールである pandas，numpy，scipy，TOT のオープンソースコード[5]用いて実装した．Intel Core i7-7820X CPU およびメモリ 64GB の計算機を利用し，OS は Ubuntu 18.04 を用いた．

実験における初期設定は以下の通りである．LDA と TOT に共通の設定として，データセットを分割期間 6 ヶ月，重なり期間 3 か月で分割した．すなわち，$p_{2i} - p_{2i-1} = 6$ヶ月であり，$p_{2i+1} - p_{2i} = 3$ヶ月を満たすように分割期間を設定した．また，トピック数$\gamma = 8$，イテレーション数$\delta = 500$とした．個別の設定として，LDA では，ハイパーパラメータα，βを gensim のデフォルト値である$1/\gamma$，$1/\gamma$とし，TOT では，α，βをオープンソースコード[5]の記述と同様に 50/γ，0.1 とした．

3 章で定義した順序付き集合族Δ_Pを LDA に入力して得られた出力を分析した結果を表 1，順序付き集合族Δ_PとT_Pを TOT に入力して得られた出力を分析した結果を表 2 に示す．各表の 1 行目は分割期間を表す．各列は実線で区切られた 8 個のトピックごとに，トピック固有の単語多項分布ϕに基づいてトピックへの所属確率が高い単語$Z = 5$個を抽出して記載している．また，抽出した単語から，ラベルとなり得る単語を調べるため，表のセルおよび文字の色付けを行った．色付け方法を以下に示す．

まず，文書固有のトピック多項分布θから各トピックへの所属確率が高い文書の上位 10 件を抽出し，文書の内容にまとまりがあるかを判別した．文書内容がまとまっているトピックは，実線で囲まれたセルを淡黄色に着色している．次に，着色したトピックに対して以下の評価を行った．上位 10 件の文書に対して，抽出した 5 個の単語がトピックのラベルとなり得るかを判別した．ラベルとなり得る場合は単語を赤色に着色している．最後に，マルウェア名や攻撃キャンペーン名はセキュリティレポートを検索する際に有用なラベルとなり得るため，赤色に着色した単語が，マルウェア名や攻撃キャンペーン名であれば，その単語に下線を引いた．LDA および TOT において，色付けされたトピックの割合を表 3 にまとめた．

表 3 より，LDA よりも TOT を用いた結果のほうがより多くのラベルとなり得る単語を抽出していることがわかる．これは，TOT が時間要素を考慮しているからだと推測できる．時間要素を考慮すると，時系列に沿ってトピックを捉えることができる．そのため，内容が類似した文書で構成されたトピックが多くなり，さらに抽出した文書とラベルとなり得る単語の対応関係が明確なトピックが多くなったと考えられる．

一方で，マルウェア名や攻撃キャンペーン名を含むトピックの割合は LDA で 32%，TOT で 33%であり，大きな差は見受けられなかった．全トピックのうちマルウェア名や攻撃キャンペーン名が含まれるトピック数は，LDA が 18 個，TOT が 22 個であり，大きな差がないことがこの要因であると考えられる．

表 3　LDA および TOT の色分けしたトピックの割合

	$\frac{\text{淡黄色のトピック数}}{\text{全トピック数}}$（%）	$\frac{\text{赤色文字のトピック数}}{\text{淡黄色のトピック数}}$（%）	$\frac{\text{下線文字のトピック数}}{\text{赤色文字のトピック数}}$（%）
LDA	50	70	32
TOT	68	86	33

5. 考察

今回の実験で得られたラベルとなり得る単語と，同時期に流行していたマルウェアや攻撃キャンペーンとの関係を検証した．まず，「meltdown」と「spectre」はともに CPU の脆弱性であり，2017 年 11 月頃に記事が公開された[6]．その後，2018 年 3 月にかけて Apple，Google，Microsoft などが Meltdown と Spectre のセキュリティアップデートを実施した．表 1 および表 2 の実験結果を見ると，それぞれ「2018/1～6（LDA）」と「2017/9～2018/3（TOT）」の期間に「meltdown」と「spectre」が出現している．TOT を用いた結果の方が，より正確に Meltdown と Spectre の流行時期を捉えていると考えられる．また，Mirai は 2016 年 11 月頃に出現した主に IoT 機器を標的としたマルウェアであり，2017 年 12 月頃から Mirai の亜種が活動している[7][8]．表 1 および表 2 の実験結果では，ともに「2017/9～2018/3（LDA，TOT）」，「2018/1～6（LDA，TOT）」，「2018/3～9（LDA，TOT）」の期間に「mirai」が出現していることがわかる．さらに、表 1 および表 2 の実験結果を見ると，「wannacry」、「petya」、「notpetya」、「bad rabbit」といったランサムウェアに関するラベルが多く出現している。WannaCry は Microsoft 製品の SMB に関連する脆弱性を利用したランサムウェアであり、2017 年 4 月にエクスプロイトコード（ETERNALBLUE）が公開され、その後、感染が急拡大した。Petya ならびに NotPetya は 2017 年 6 月下旬に欧州を中心に拡散したランサムウェアであり、ウクライナ製の会計ソフトの MeDoc のアップデートパッケージを介して感染が拡大した[9]。BadRabbit は 2017 年 10 月から 11 月にかけて拡散したランサムウェアであり、Petya（NotPetya）の亜種と指摘されている。ここで、表 1 および表 2 の実験結果を見ると，上記のランサムウェアに関する出力結果の出現時期はともに 2017 年全般であるが、出現回数は TOT を用いた結果の方が多く、「bad rabbit」は TOT を用いた結果のみに出現している。よって、TOT を用いた結果の方が，より正確に上記のランサムウェアの流行時期を捉えていると考えられる．さらに、提案手法では「bad rabbit」のように、ある時期に

局所的に出現したマルウェア名をラベルとして捉えることができているが、これらのラベルはデータセットを一括で入力した場合には適切に出力されないことが考えられる。以上の例のように，表 1 および表 2 の実験結果に登場したラベルとなり得る単語は，同時期に流行していたマルウェアや攻撃キャンペーンを捉えていると考えられる．

マルウェアや攻撃キャンペーン以外にもラベルとなり得る単語は多く見受けられた．表 1 および表 2 の実験結果では，ともに「2018/3～9（LDA, TOT）」と「2018/6～12（LDA, TOT）」の期間に「blockchain」や「transaction」，「miner」などの単語が出現している．これは，2018 年初頭にビットコインの価格が暴落したことを受けて，ブロックチェーン技術の安全性や実現可能性に関する記事が多く発行されたからである．また，表 1 および表 2 の「2017/1～6」~「2018/1～6」の期間に「healthcare industry」や「healthcare organization」，「medical device」といった医療業界と関連した単語が出現している．医療業界のサイバー攻撃被害数は 2016 年ごろから増加しており，2017 年には，サイバー攻撃のおよそ半分以上が医療機関を標的としたものであった[10]．この原因は，医療業界が患者へのケアや医療技術の進歩に重点を置くあまり，他業界に比べてサイバーセキュリティ対策への費用が十分でないことを背景にしている．そのため，医療業界はランサムウェアなどの標的となりやすい傾向にあり，医療業界の被害状況や，医療機関へ事前対策および事後対応を促す記事が多く発行された．これら以外にも，「smb」や「Android」などラベルとなり得る単語は多く存在した．以上より，提案手法により得られた結果は，常用的なフレーズではなく，ラベルとして有用な単語を多く含んでいると考えられる．

6. まとめ

本稿では，トピックモデルを利用してセキュリティレポートから特徴的なラベルを抽出する手法を提案した．そのアイデアは，データセットを直接トピックモデルに入力するのではなく，重なり期間のある一定期間ごとに分割されたデータセットを入力することである．提案方法を用いることにより，ラベルに適した単語を含むトピックを多く出力することができた．また，出現頻度の高い一般的な単語を含むトピックを削減できることも示唆された．

実験では 2400 件程度の文書を入力として使用し，LDA ならびに TOT を利用した場合のラベルの出力結果を検証した．考察で述べたように，流行時期を捉えたマルウェアや攻撃キャンペーンに関するラベルやその他のラベルとなり得るフレーズが出力されていることを確認できた．今後の研究の方向性として，レポート件数が増えて数万件規模になったときに，提案手法のパラメータとトピックモデルのハイパーパラメータの調整により同様に動作するかを検証する予定である．また，本稿の実験では提案手法のパラメータを固定していたが，パラメータを変更してラベルを生成すれば，同一文書に異なるラベルが付与されるので，検索性能を向上させる異なるラベルの利用方法についても検討する予定である．

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] David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, no.3, pp.993-1022, 2003.
[2] Xuerui Wang, Andrew McCallum. 2006 “Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends” In ACM SIGKDD
[3] “Natural Language Toolkit” https://www.nltk.org/
[4] “genism: models.ldamodel – Latent Dirichlet Allocation” https://radimrehurek.com/gensim/models/ldamodel.html
[5] “Topics Over Time” https://github.com/ahmaurya/topics_over_time
[6] “Kernel index [LWN.net] - Meltdown and Spectre” https://lwn.net/Kernel/Index/#Security-Meltdown_and_Spectre
[7] “Rise of One More Mirai Worm Variant” https://www.fortinet.com/blog/threat-research/rise-of-one-more-mirai-worm-variant
[8] “Warning: Satori, a Mirai Branch Is Spreading in Worm Style on Port 37215 and 52869” http://blog.netlab.360.com/warning-satori-a-new-mirai-variant-is-spreading-in-worm-style-on-port-37215-and-52869-en/
[9] “New ransomware, old techniques: Petya adds worm capabilities” https://www.microsoft.com/security/blog/2017/06/27/new-ransomware-old-techniques-petya-adds-worm-capabilities/?source=mmpc
[10] “Top Cyber Trends in Healthcare Today” https://www.fortinet.com/blog/business-and-technology/top-cyber-trends-in-healthcare-today--q-a-with-sonia-arista--nat

表 1　分割データセットを LDA に入力した結果

2017/1-6	2017/3-9	2017/6-12	2017/9-2018/3	2018/1-6	2018/3-9	2018/6-12	2018/9-2019/3	2019/1-6	2019/3-9	2019/6-12
wannacry	wannacry	petya	wannacry	blockchain	wannacry	wannacry	federal agency	msps	persistence	dns
keen lab	shellcode	wannacry	casb	hash	msps	splunk	atm	accuracy	threat detection	prediction
team sniper	exploit kit	nist csf	dns	wannacry	ios device	api	scammer	fraud	cwp	security analytics
microsoft edge	python	ukraine	ise	apple	salesforce	msps	cdm	rsa	dataset	five pillar
tencent security	worm	equifax	google play	disclose	mitigation	samsam	api	cio	medr	holiday season
shamoon	msp	asaf	federal agency	vms	vpnfilter	blockchain	magecart	scammer	red team	scammer
endpoint protection	security vendor	soc	temp	dns	taa	transaction	apple	certificate	medical device	potential victim
certificate	msp program	anomaly	ise	rdp	china	small business	ise	five pillar	mfa	response team
middle east	msps	dmarc	vms	macro	dlp	apple	installation	active directory	steal data	workforce
credential theft	trump administration	stanford	macro	miner	rdp	medical device	british airway	casb	ddos	malicious apps
botnet	petya	bot	shopper	casb	carrier	msps	iot device	com	casb	local government
india	medical device	iot device	onedrive	cloud apps	casb	federal agency	rdp	dot	xdr	dataset
hitrust	wmi	default password	holiday season	google play	sep	classifier	bot	infected computer	siem	threat detection
online safety	ukraine	debugger	bot	response team	overview	byte	umbrella	china	accuracy	xdr
local government	wikileaks	com	healthcare industry	web isolation	mssp	phone number	contractor	eternalblue	dlp	municipality
dpa	nist csf	cryptocurrency	dmarc	smbs	blockchain	atm	taa	gmail	msps	shellcode
csf	healthcare organization	shopper	nist csf	appdata	transaction	carrier	threat grid	threat response	tweet	registry
tcp	healthcare industry	notpetya	healthcare organization	folder	hash	exploit kit	com	threat grid	dns	persistence
exploit kit	respondent	local government healthcare industry	controller	submission	miner	rdp	middle east	apple	utility	volatility
python	authority		sender	saas	cryptocurrency mining	obfuscation	victim environment	scammer	backup system	dll
wmi	endpoint protection	ascii	scammer	mirai	iot device	ise	classifier	honeypot	scammer	casb
apple	ciso	amp	iot device	scammer	mirai	iot device	threat hunting	persistent threat	dot	workplace
worm	impersonation	macro	mirai	ciso	vmware	google play	msps	iot device	potential victim	workforce
confidential data	audit	indicator	bot	ios device	simulation	ngfw	byte	theft	malware analysis	unmanaged device
proxy	packet	com	default password	iot device	ciso	pxgrid	account takeover	compromised account	response team	duo
shellcode	keen lab	google play	threat grid	meltdown	macro	vmware	amazon	red team	rsac	msps
authority	team sniper	ise	packet	spectre	dns	iot device	smb	registry	threat hunting	xdr
conference	tencent security	social medium android device msp program	behavioral indicator	iot device	istr	taa	vmware	com	vast majority	siem
ise	uaf		spectre	devsecops	temp	umbrella	active directory	istr	email threat	cybersecurity team
smb	microsoft edge		meltdown	transparency	following command	grid	admin	scheduled task	prevention	binary
medical device	amp	dmarc	sep	sep	binary	dlp	small business	rdp	cybersecurity team	backup system
cerber	apple	encrypted traffic	cio	binary	appdata	node	registry	smb	cio	dataset
nist csf	nonprofit	controller	stakeholder	iiot	folder	smart device	iiot	mitre att	board member	utility
healthcare organization	micro	notpetya	shortage	com	apple	iot device	rsac	foothold	metric	directory
csf	tcp	amp	ciso	byod	app store	smbs	infected computer	iot device	gmail	investigator
macro	macro	respondent	equifax	federal agency	devsecops	threat grid	exploit kit	source code	source code	threat hunting
amp	tor	linux	folder	threat grid	threat grid	loader	ai model	carbanak	carbanak	medr
com	redirect	tor	transparency	cio	devops	triton	scammer	dataset	istr	utility
redirect	cerber	exploit kit	disclose	healthcare industry	response team	magecart	emea	federal agency	persona	banking trojan
spam	malicious document	smbs	transparent	kevin simzer	google play	middle east	governance	binary	tweet	mobile security

表 2　分割データセットを TOT に入力した結果

2017/1-6	2017/3-9	2017/6-12	2017/9-2018/3	2018/1-6	2018/3-9	2018/6-12	2018/9-2019/3	2019/1-6	2019/3-9	2019/6-12
wannacry	wannacry	wannacry	wannacry	ciso	wannacry	wannacry	iot device	rdp	tweet	scammer
exploit kit	macro	petya	smbs	wannacry	iot device	dlp	rdp	casb	iran	webinar
botnet	tcp	notpetya	atm	rdp	backup	healthcare	com	dataset	casb	potential victim
cerber	security vendor	ukraine	worm	com	carrier	samsam	smb	certificate	com	vpn
com	wannacry ransomware	cryptocurrency	petya	stakeholder	healthcare	endpoint protection	classifier	malware analysis	dataset	ddos
shamoon	nist csf	linux	bot	backup	recipient	carrier	facebook	scammer	red team	threat detection
shellcode	iot device	conference	directory	vms	ise	atm	ise	apple	mfa	dlp
hitrust	shamoon	proxy	macro	healthcare industry	ciso	small business	theft	login credential	accuracy	endpoint detection
middle east	trump administration	social medium	folder	healthcare	simulation	scammer	api	utility	provider	saas
python	healthcare organization	packet	middle east	iiot	enterprise network	exploit kit	cisos	undetected	tweet	medr
wannacry	shellcode	iot device	scammer	bitcoin	workload	byte	saas	devops	msps	dataset
petya	wmi	default password	casb	soc	federal agency	federal agency	amazon	cybersecurity team	gartner	workforce
fbi	api	console	google play	controller	conference	devops	casb	medical device	amp	casb
smb	petya	workplace	cloud apps	bot	transparency	cyber	vmware	disclose	threat hunting	apis
ukraine	powershell	atm	webinar	trojan	perimeter	classifier	cwp	rsa	threat detection	prediction
certificate	linux	ascii	bitcoin	threat grid	android	bec	federal agency	red team	botnet	xdr
dlp	tor	com	vms	appdata	dlp	cryptocurrency	msps	accuracy	cio	backup system
online safety	ukraine	macro	default password	casb	phone number	blockchain	devops	istr	metric	siem
proxy	petya	security vendor	transparent	macro	scammer	transaction	small business	mitre att	administrator	prevention
confidential data	cloud ready	financial institution	byte	sep	casb	android	governance	installation	protocol	volatility
medical device	amp	respondent	equifax	dns	blockchain	iot device	payload	active directory	cybersecurity team	binary
apple	bot	ise	sep	amp	transaction	china	china	china	webinar	directory
ciso	redirect	nist csf	dns	recipient	hash	bot	registry	rto	devops	byte
trump administration	android device	healthcare organization	federal agency	ddos attack	bitcoin	com	ai model	threat hunting	backup system	query
wmi	botnet	webinar	apple	china	miner	simulation	taa	iocs	cwp	signature
keen lab	keen lab	equifax	iot device	mirai	mirai	msps	scammer	com	intrusion	dns
team sniper	team sniper	google play	nist csf	ise	consent	umbrella	soc	registry	powershell	msps
tencent security	apple	cloudsoc	healthcare organization	iot device	vpnfilter	cisos	rsa conference	federal agency	utility	cybersecurity team
uaf	tencent security	worm	mirai	container	smbs	google play	iiot	msps	ttps	cloud provider
bot	uaf	amp	packet	devsecops	container	attendee	emea	smb	default	meantime
conference	cerber	exploit kit	controller	apple	threat grid	ise	amp	iot device	scammer	shellcode
india	social medium	tor	meltdown	folder	powershell	vmware	threat grid	fraud	siem	linux
threat grid	binary	prediction	spectre	scammer	russia	cwp	indicator	persistent threat	xdr	source code
amp	germany	kit	bad rabbit	meltdown	com	api	magecart	bot	leverage	mobile security
api	pawn storm	integrity	shopper	google play	binary	binary	threat hunting	accountable	recipient	google play
dpa	medical device	dmarc	dmarc	federal agency	folder	soc	atm	source code	source code	utility
csf	respondent	sender	threat grid	cloud apps	appdata	protocol	ransom	carbanak	dns	disaster recovery
macro	endpoint protection	smbs	sender	smbs	apple	taa	cryptocurrency	cio	carbanak	ise
tcp	nonprofit	asaf	dashboard	encrypted traffic	javascript	loader	samsam	metric	android	persistence
nist csf	healthcare industry	impersonation	behavioral indicator	workload	macos	triton	transaction	ctr	executable	tco