Vol.2020-SPT-36 No.45
2020/3/3

セキュリティレポートの時系列トピックモデルを用いた分析

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：Topics Over Time（TOT）は時間とともに変遷するトピックをとらえるためのトピックモデルである．セキュリティレポートに TOT を適用することで，マルウェアやサイバー攻撃などの時間的変遷を分析できると期待される．2017 年から 2019 年のセキュリティレポートに TOT を適用した結果，特徴的な変遷を捉えられることが示唆された．

キーワード：トレンド分析，Topics Over Time，トピック分布，単語分布，セキュリティレポート

Analysis of Security Report Using Continuous-Time Topic Models

NAGASAWA RYUSEI[†1]　FURUMOTO KEISUKE[†2]　TAKITA MAKOTO[†3]
SHIRAISHI YOSHIAKI[†1]　TAKAHASHI TAKESHI[†2]　MOHRI MASAMI[†4]
TAKANO YASUHIRO[†1]　MORII MASAKATU[†1]

Abstract: Topics Over Time (TOT) is a topic model to be aware of topics that change over time. By applying TOT to security reports, it is expected that transition of malware and cyber-attacks can be analyzed. As our case study of applying TOT to security reports from 2017 to 2019, it is suggested that characteristic changes could be captured.

Keywords: Trend Analysis，Topics Over Time，Topic Distribution，Word Distribution，Security Report

1. はじめに

大規模かつ不均質な大量のテキスト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る．トピックモデルでは，文書ごとのトピックの構成比率であるトピック分布と，トピックごとの単語の比率である単語分布を推定する．代表的なものに LDA（Latent Dirichlet Allocation）[1]などがある．ニュース記事やブログ記事，論文などのトピックの発生は，時間とともに変化する．しかし，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．

トレンド分析がマーケティングや為替など様々な分野に適用されている．トレンド分析とは，データの中から時系列を考慮してトピックを選出し，トピックの変遷を分析することである．時系列トピックモデル TOT（Topics Over Time）[2]が Wang らによって提案されている．

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．その内容は時々刻々と変化する．そこで，セキュリティレポートに TOT を適用することで, マルウェアやサイバー攻撃などの時間的変遷を捉えることができるか確認する．

2. 時系列トピックモデル

2.1 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 との違いは，トピックを推定する際に単語の文書ごとの共起情報だけではなく，時間情報を考慮することである. すなわち TOT では文書中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する．TOT の文章生成過程であるグラフィカルモデルを図 1 に示す.

TOT はある文章にあるトピックが現れる確率 θ，あるトピックにある単語が現れる確率 ϕ とともに，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．つまり，入力文書集合から図 1 の θ，ϕ，ψ を出力する．入力は, 文章ごとの単語集合と, 文章ごとの時間集合である.

2.2 適用例

セキュリティレポートに TOT を適用する準備として，PNAS（国立科学アカデミー論文集）の 1915 年から 2005 年の論文のタイトル[3]を用いて予備実験をした．データセットの文書数は 79801 件である．

前処理として, a や when などのストップワードの除去の

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

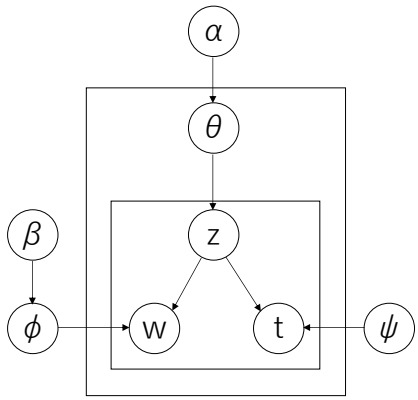

図 1　Topics Over Time (TOT)のグラフィカルモデル

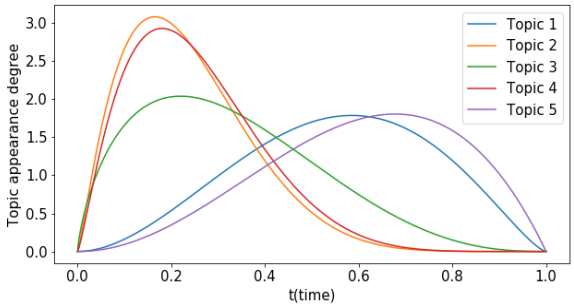

図 2　PNAS におけるトピックの時間遷移

みを行った．トピック数は T=5 とし，ディリクレ事前分布のパラメータ α，β はそれぞれ 50/T, 0.1 としたときの TOT の適用結果を図 2，図 3 に示す．図 2 はトピックの時間遷移を表しており，垂直軸がトピックの出現度合い，水平軸が[0,1]に正規化した時間を表す．図 3 は，トピックに対して関係の深い単語の分布である．ここで，関係の深い単語とは，あるトピックにある単語が現れる確率 ϕ が閾値9×10^{-3}以上（文献[3]と同一の値）の単語を指す．

図 2 より時間経過に伴うトピックの推移を確認できる．また図 3 より，それぞれのトピックにはそのトピックを特徴づける単語が分布していることがわかる．トピック内の単語を調べることで，そのトピックがどのような話題をもとに作られたのか解釈する．

3. セキュリティレポートへの TOT の適用

3.1 データセット

データセットは，セキュリティベンダー8 社（Arbor[4], Barracuda[5], Cisco[6], Druva[7], FireEye[8], Paloalto[9], Symantec[10], TrendMicro[11]）のブログページから，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で 2386 件からなる．

3.2 前処理

2.2 節の PNAS データセットに対しては，ストップワードの除去のみを行って TOT に適用した．しかし，セキュリティレポートの本文を入力するときには，単語集合に数字

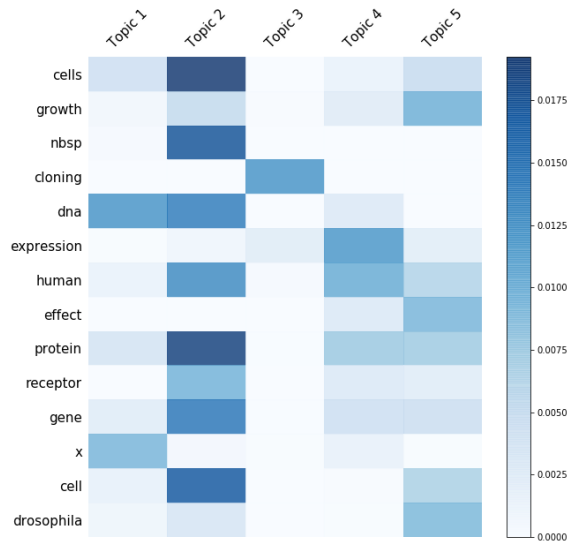

図 3　PNAS におけるトピックに対する関係の深い単語の分布

や記号が混ざった単語や，出現回数が極めて少ない単語など，入力に適さない単語が多く含まれている．また，複数の単語を組み合わせて意味を成す複合語も多く存在する．そこでセキュリティレポートのデータセットでは，以下のような処理を行い，文書ごとの単語集合を作成した．

・複合語の生成
・ストップワードの除去
・数字，記号，引用符などが含まれる単語の除去
・製品名の除去
・出現回数が一回の単語を除去

3.3 予備実験と結果

前処理後の単語集合と文書の時間集合を入力として TOT を適用した．2.2 節と同様にトピック数 T=5 とし，ディリクレ事前分布のパラメータ α，β はそれぞれ 50/T，0.1 とした．その結果を図 4 に示す．

図 4 ではすべてのトピック出現時期のピークが真ん中に集まっている．PNAS データセットは学術論文の集まりなので，使用される単語が一定の範囲に収まっている．また，2.2 節では論文の本文ではなくタイトルを入力としているので，出現する単語の総数が少ない．具体的には一タイトルを前処理すると，高々15 単語ほどしか単語集合に残らない．これらにより，トピック数 T が少なくても，それぞれのトピックに明確な差異を生むことができたと考えられる．

一方，セキュリティレポートは，ある特定の話題が同工異曲で書かれている，すなわち，同じ話題でも著者によって使用される単語が異なることがある．これにより，同意義の単語が乱立しやすくなる．また，TOT への入力をセキュリティレポートのタイトルではなく本文を用いているため，一文書を前処理した後でも数十から数百の単語が残る．

Vol.2020-SPT-36 No.45
2020/3/3

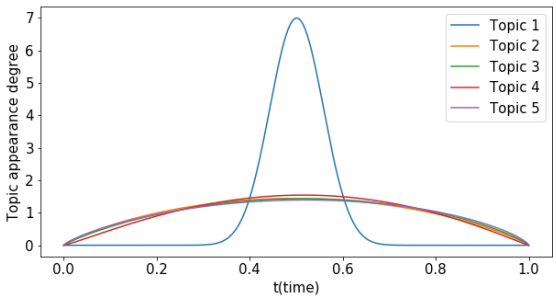

図 4　セキュリティレポートにおけるトピックの時間遷移（T=5，α=50/T，β=0.1）

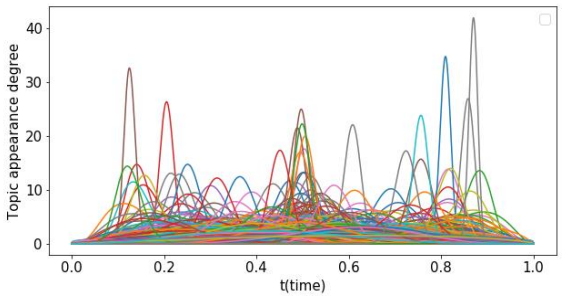

図 5　セキュリティレポートにおけるトピックの時間遷移（T=1000，α=50/T，β=0.1）

そのため，トピック数 T が小さいと各トピックが単語の特徴をとらえることができず，図 4 のような結果になったと考えられる．

3.4 TOT を適用するための方法

3.3 節で示した原因に対して以下の二点を行う．

- 一文書を前処理して生成した単語集合から，ある一定の割合でランダムに単語を取り出す．これを文書数だけ繰り返す．
- トピックごとの特徴的な分布が現れるまで，トピック数を増やす．

なお以下の実験では 20%の単語をランダムに選んでデータセットを構築した．前者の単語集合の総数を減らすことで，不要な単語の乱立を抑制できると期待される．また，後者のトピック数を多くしていくことで，一つのトピックが包括する単語の数が減り，特徴的なトピックが生まれやすくなると期待される．これら二点の処理を加えて TOT を適用したところ，トピック数 T=1000 のとき，特徴的なトピックを多く含む時間遷移が得られた．その結果を図 5 に示す．

図 5 より，図 4 とは異なりそれぞれのトピックが特徴的に遷移していることがわかる．

4. ケーススタディ

4.1 特徴的なトピック

1000 個のトピックの中から，特徴的な 3 つのトピックを抽出し，考察する．

図 6，図 7，図 8 において，3 つともに局所的なトピックの出現を示している．トピック 70 では，2017 年の 3 月から 2018 年の 6 月頃，トピック 121 では，2018 年の 2 月から 8 月頃，トピック 47 では，2018 年の 10 月から 2019 年の終わり頃にそれぞれトピックが集中して分布している．

図の下に，トピックと関係が深い文書番号を示した．これは，ある文章にあるトピックが現れる確率 θ を用いて導き出したものである．セキュリティレポートの文書数は 2386 件からなり，出版年で昇順にソートされている．つまり，文章番号をみてトピックの偏りを調べられる．

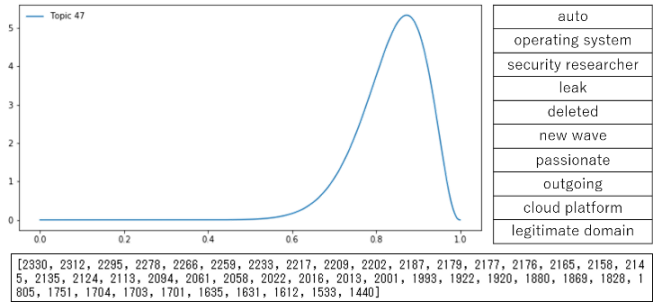

図 6　トピック 47 の時間遷移（左上），トピックに関連の深い単語（右上），トピックに関連の深い文書番号（下）．

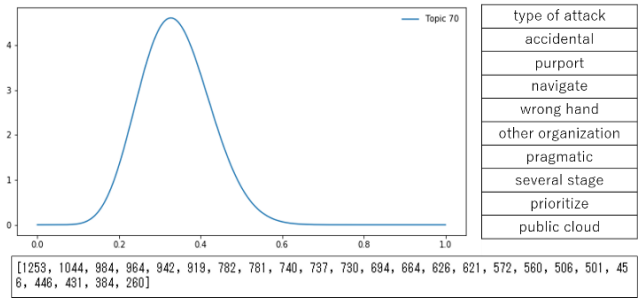

図 7　トピック 70 の時間遷移（左上），トピックに関連の深い単語（右上），トピックに関連の深い文書番号（下）．

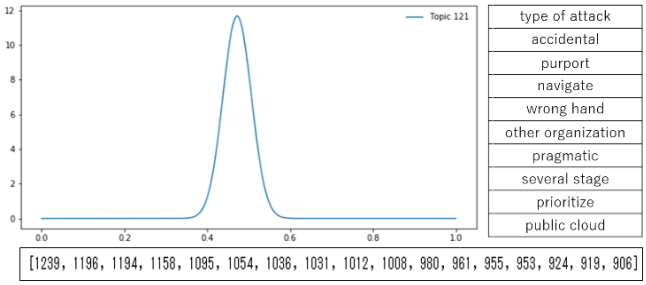

図 8　トピック 121 の時間遷移（左上），トピックに関連の深い単語（右上），トピックに関連の深い文書番号（下）．

図 6，図 7，図 8 を見ると，トピック 47 では，番号が大きい文書が固まっている．また，トピック 70 では，番号の小さい文書が集中している．さらに，トピック 121 では文

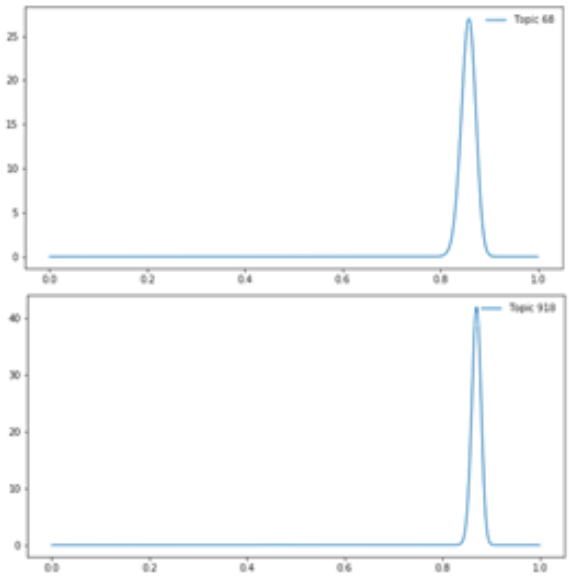

図 9　ピークが近いトピック（トピック 68，トピック 918）

表 1　分布が近いトピックの類似度

	単語数	文書数	両方のトピックに含まれる単語数	両方のトピックに含まれる文書数
トピック 68	187	60	31	15
トピック 918	304	58		

書番号が中央付近に固まっている．この結果より，どのトピックも時間遷移図と関係の深い文書が対応していることがわかる．

次に，3 つのトピックの単語の時系列分布を調べる．図の右上にトピックと関係が深い単語上位 10 個を列挙した．トピックに関係が深い単語は，あるトピックにある単語が現れる確率 ϕ をもとに算出した．ここで単語の時系列分布を調べる目的としては，文書番号のようにトピック発生がピークとなる時期に単語が集中しているのかを確認するためである．

トピック 47 は，“cloud platform”，“legitimate concern”の二単語はトピックのピーク周辺に分布していた．しかし，ほかの単語はほとんど一様分布しているものが多かった．トピック 70 もトピック 47 と同様に，“wrong hand”，“several stage”の二単語はトピックのピーク周辺に分布していたが，ほかの単語はほとんど一様分布していた．トピック 121 は，“false sense of security”の一単語のみトピックのピーク周辺に分布していた．この結果から，トピックの偏りは，ϕ から算出された単語すべてが影響しているわけではないことがわかる．

時間遷移図において，分布が近いトピック，遠いトピックに注目する．トピック 68 とトピック 918 は図 9 をみると，ほぼ同じ時期にピークを迎えていることがわかる．この 2 つのトピックは，時間遷移図では似たような発生分布を示しているが，トピックと関係の深い単語や文書に着目

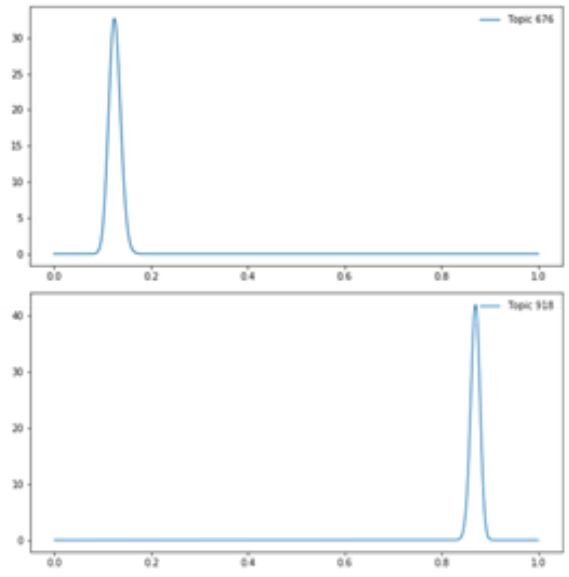

図 10　ピークが遠いトピック（トピック 676，トピック 918）

表 2　ピークが遠いトピックの類似度

	単語数	文書数	両方のトピックに含まれる単語数	両方のトピックに含まれる文書数
トピック 676	159	38	22	0
トピック 918	304	58		

したとき，その類似度はトピック分布の近さによって変わるのか確認する．トピック 68，トピック 918 の類似度を見るため，それぞれに含まれる単語数，文書数，両方のトピックに含まれる単語数，文書数を表 1 に示す．表 1 より，トピック 918 に含まれる単語の約 10%がトピック 68 と同じ単語であり，約 26%の文書が同じであることがわかった．

時期的に分布が遠い図 10 に示すトピックについても確認する．トピック 676 とトピック 918 は発生時期のピークが全く異なる．それぞれに含まれる単語数，文書数の両方のトピックに含まれる単語数，文章数を表 2 に示した．表 2 より，トピック 918 に含まれる単語の約 7%がトピック 68 と同じ単語であり，文書に関しては 2 つのトピックに共通して含まれる文章は存在しなかった．表 1 と表 2 の結果から，トピック発生の時期的な近さは，トピックの内容に関係があることがわかる．

5. まとめ

セキュリティレポートに TOT を適用することで，時間要素を踏まえたうえで特徴的なトピックを出力できることが示唆された．

セキュリティレポートの本文は前処理の方法によっては似たような単語が複数の文書に含まれることになり，トピック間の差異を見出しづらくなることがある．TOT の適用結果に応じてどのトピックにも広く分布している単語を除去し，トピックを特徴づける単語が埋没しにくくなるよ

うにする方法を与えることが今後の課題としてあげられる.

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. 2003 “ Latent Dirichlet Allocation ” Journal of Machine Learning Research, no.3, pp.993-1022
[2] Xuerui Wang, Andrew McCallum. 2006 “ Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends ” In ACM SIGKDD
[3]“Topics Over Time” https://github.com/ahmaurya/topics_over_time
[4] “NETSCOUT’s ASERT Blog”https://www.netscout.com/asert
[5] “the Barracuda blog” https://blog.barracuda.com/
[6] “Security Archives – Cisco Blogs” https://blogs.cisco.com/security
[7] “Druva Blog – Data Protection for the Cloud Era | Druva” https://www.druva.com/category/tech-engineering/
[8] “Threat Research | FireEye lnc” https://www.fireeye.com/blog/threat-research.html
[9] “Palo Alto Networks – Unit42” https://unit42.paloaltonetworks.com/?pg=1#threat-brief
[10] “Symantec blogs” https://www.symantec.com/blogs/
[11] “Simply Security News, Views and Opinions from Trend Micro” https://blog.trendmicro.com/