Computer Security Symposium 2017
23 - 25 October 2017

メールとそのコンテキスト情報を基にした不審メール検知手法の提案

西川 弘毅[†1]　山本 匠[†1]　河内 清人[†1]

概要：特定の組織や人を対象として、機密情報の窃取等の攻撃を行う標的型攻撃は深刻な脅威となっている。攻撃の起点の多くは、標的型メールとなっている。そのため、標的型攻撃メールを防ぐことは、巧妙となるサイバー攻撃を防ぐ観点から重要である。しかし、既存の技術では、巧妙な攻撃者による標的型攻撃メールを検知することができない。具体的には、攻撃者が、既に標的と関係がある組織の踏み台に感染した状態で、最終目標となる対象への感染を目的とする場合に、踏み台のメールアドレスや情報を利用して最終目標にメールを送ることが考えられる。この場合には、踏み台の特徴を踏まえたうえで攻撃のメールを送るため、既存の技術では検知することが困難である。本書では、メールのコンテキストと添付ファイルや URL 先のウェブページといったコンテンツのコンテキストとを組み合わせることで、巧妙な攻撃者による攻撃を検知する技術を提案する。

キーワード：　不審メール検知、標的型攻撃

Detection method what suspicious email is based on email context

Hiroki Nishikawa[†1]　Takumi Yamamoto[†1]　Kiyoto Kawauchi[†1]

Abstract: Targeted attack is serious problem for stealing confidential information. Especially, beginning of targeted attack is mainly spear phishing email. Therefore, detection of spear phishing email is important for protecting from sophisticated attack. However, previous method can't detect sophisticated attack that attacker sends finally target when attacker already infected springboard related the target. In this case, it is difficult to detect the email using previous method because attacker make email considering feature of springboard. In this paper, we propose detection method what suspicious email is based on email context.

Keywords: Suspicious email detection, Targeted attack

1. はじめに

特定の組織や人を対象として、機密情報の窃取等の攻撃を行う標的型攻撃は深刻な脅威となっている。標的型攻撃の起点は、対象組織に特化した内容を送る標的型メールが主流である。トレンドマイクロの調査では、この標的型攻撃メールによるマルウェア感染は、企業に対する攻撃全体の 76%にも上るとの結果が出ている[1]。そのため、標的型攻撃メールを防ぐことは、巧妙となるサイバー攻撃を防ぐ観点から重要である。

不審メールを検知する技術として、受信メールの送信ドメイン認証結果や送信経路、添付ファイルの名称やアイコン偽装をもとに不審メールを判断する技術[2]や、添付ファイルのマクロが悪性である可能性がある場合には、マクロを削除したドキュメントを再構築することで無害化する技術[3]や、識別対象のメールを収集し、特徴量を SVM で学習、学習した分類器により、受信したメールが、予め学習した人物からのものであるかを判定する技術[4]がある。しかし、これらの技術では、巧妙な攻撃者による標的型攻撃メールを検知することができない。具体的には、攻撃者が、既に標的と関係がある組織の踏み台に感染した状態で、最終目標となる対象への感染を目的とする場合に、踏み台のメールアドレスや情報を利用して最終目標にメールを送ることが考えられる。この場合には、踏み台の特徴を踏まえたうえで攻撃のメールを送るため、既存の技術では検知することが困難である。

しかし、攻撃者が踏み台を通じてメールを送る際に、添付ファイルや URL 先のウェブページといったコンテンツの情報を、攻撃者自らが容易したコンテンツに差し替えて、メールの趣旨に添わないもので送る可能性がある。その場合、メールの題名、宛先、本文等から、メールのコンテンツが不自然な状態であると推測できる。そのため、正常な状態と、攻撃者によりメールのコンテンツが差し替えられている状態では、メールの題名、宛先、本文等の情報をもとにして得られる特徴（以下、メールのコンテキストと呼ぶ）と、添付ファイルや URL 先のウェブページといったコンテンツの情報をもとにして得られる特徴（以下、コンテンツのコンテキストと呼ぶ）との間の関係性に差異が生じると考えられる。そこで、まず、正常な状態におけるメールのコンテキストとコンテンツのコンテキストとの関係性を学習しておく。その後、検査対象となるメールからメールのコンテキストとコンテンツのコンテキストを抽出し、検査対象メールのコンテキストを、学習済みのメールのコ

†1 三菱電機株式会社　情報技術総合研究所
Mitsubishi Electric Corporation, Information Technology R&D Center.

ンテキストとコンテンツのコンテキストとの関係性に照らし合わせることで、本来得られるコンテンツのコンテキストを取得し、検査対象メールのコンテンツのコンテキストとを比較した際に、類似度が閾値より低い場合は、コンテンツが差し替えられている可能性が高いとし、不審と判断する。

本書の構成は、2 章で関連研究について示し、3 章では今回の手法で対象とする攻撃の範囲について示す。4 章では提案する検知手法について説明する。5 章では、提案手法の課題等について考察する。

2. 関連研究

本章では、提案手法に関連する、不審メール検知技術について説明する。

CipherCraft/Mail[2]は、受信メールを、送信ドメイン認証結果や送信経路といった挙動と、名称やアイコン偽装といった添付ファイルに関する不審点をもとに検査し、自動隔離・注意喚起する技術である。しかし、信頼のおける人物に感染した後に、その人物のメールアドレスを利用してメールを送る攻撃では、挙動に関する不審点は検知できず、高度な攻撃者による添付ファイルが作成される場合、サンドボックスによる検知を通過するため、本技術では検知できない

Disarm[3]は、添付ファイルのドキュメントが悪性である可能性があるコード（マクロ等）を含む場合、該当コードを除去し、ドキュメントを再構成することで、悪性マクロの実行を予防する。しかし、マクロ等を活用している組織である場合、Disarm を無効にすることが公式で推奨されているため、そのような組織では有効に働かない。

Sevtap D らの手法[4]は、識別対象のメールを収集し、特徴量を SVM で学習、学習した分類器により、受信したメールが、予め学習した人物からのものであるかを判定する技術を提案している。しかし、認識精度は 67%～100%とまばらであり、確度を持って本人からメールであると言うには信頼性が低いことと、本人識別を通過するように、本人の特徴を学習する巧妙な攻撃には無力である課題がある。

3. 検知対象

本手法により検知する対象の攻撃について説明する。
攻撃者は、予め、最終目標の標的に感染するために、踏み台となる組織の端末に感染しているものとする。踏み台となる組織は、標的組織と取引があり、かつ標的組織よりセキュリティポリシーが緩い企業である。
攻撃者が踏み台を経由して送られる標的型メールは、攻撃用メールコンテンツによって次の四つに分類することができる。

① メールコンテンツの内容が、メール本文から逸脱しており、アンチウイルスソフトにより検知可能な攻撃
② メールコンテンツの内容が、メール本文の内容と合致しており、アンチウイルスソフトにより検知可能な攻撃
③ メールコンテンツの内容が、メール本文から逸脱しており、アンチウイルスソフトで検知できない攻撃
④ メールコンテンツの内容が、メール本文の内容と合致しており、アンチウイルスソフトで検知できない攻撃

これらの攻撃の種類に対して、本手法では、③の攻撃をスコープとしている。

4. 提案手法

4.1 全体概要

本手法では、メールのコンテキストと、メールに付されたコンテンツのコンテキストに相関があることに着目した。正常なメールのコンテキストと、メールに付されたコンテンツとの関係性を学習することで、メールの文面は似せられているが、送信されているコンテンツに相関が無いようなメールを検知することができるようになる。

本手法の全体の流れを説明する。

本手法は大きく、準備段階と、運用段階、の二段階に分かれる。まず、準備段階では、学習対象となるメール 1 つ以上を含むメール集合から、メールとコンテンツとの関係性を学習する。次に、運用段階では、不審であるかを判定する対象の受信メールに対して、そのメールと、コンテンツとの関係性を、既に登録されている関係性と比較することで、受信したメールが不審かどうかを判定する。準備段階と運用段階それぞれの流れ図を図 1、図 2 に示す。

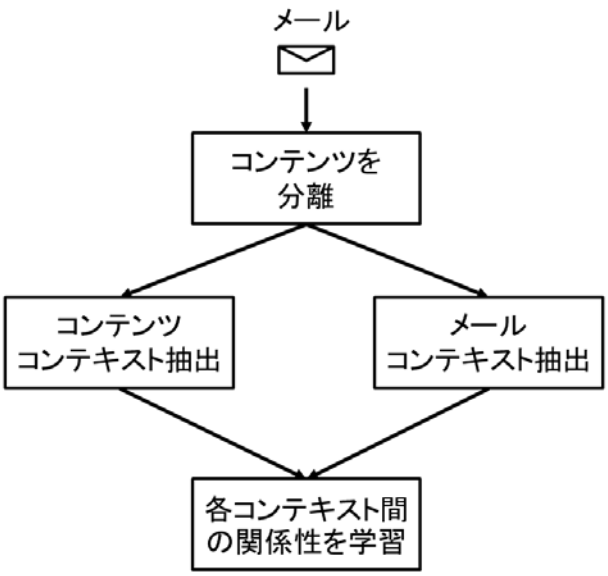

図 1 準備段階の処理全体像

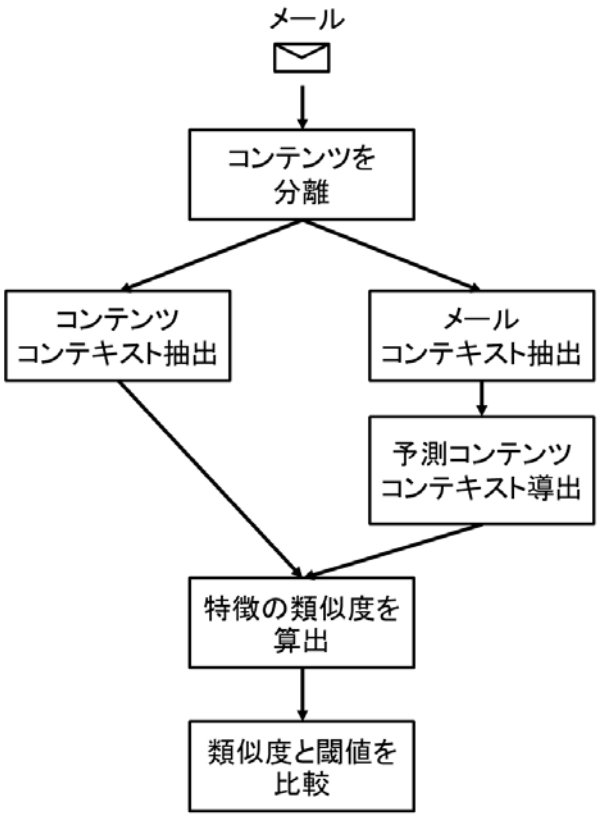

図 2 運用段階の処理全体像

続いて、各段階の処理の詳細について説明する。

まず、準備段階の処理について説明する。準備段階のフローチャートを図 3 に示す。

処理の前に、まず、学習対象となる 1 つ以上のメール集合を用意する。この時、メール集合は全て、コンテンツを含んでいるものとする。

① メール集合を、キー情報によりラベル付していく。ここでは、宛先情報をキー情報とし、集合が分類されるものとする。このラベル付を、学習対象となるメール集合が空になるまで行う。

② キー情報ごとに分類した各メール集合からメールを取り出し、メールからコンテンツを分離する。

③ メールの題名や、To, CC、本文をもとに、メールのコンテキストが抽出できる形に整形する。整形後のメールデータは、例えば、題名、宛先情報、本文、の三つの要素からなる。ここで、本文は、元々の文章から引用文や署名などを取り除き、解析しやすい状態に修正しておく。

④ 整形後のメールデータとコンテンツのそれぞれから、コンテキストを抽出する。整形後のメールデータやコンテンツからのコンテキスト抽出については 4.2 節に示す。

⑤ メールコンテキストからコンテンツコンテキストが導かれる関数を導出する。

次に、運用段階の処理について説明する。運用段階のフローチャートを図 4 に示す。

① 不審であるかを判定する対象のメールからコンテンツを分離する。

② メールの題名や、To, CC、本文をもとに、メールのコンテキストが抽出できる形に整形する。整形後のメールデータは、例えば、題名、宛先情報、本文、の三つの要素からなる。ここで、本文は、元々の文章から引用文や署名などを取り除き、解析しやすい状態に修正しておく。

③ 整形後のメールデータとコンテンツのそれぞれから、コンテキストを抽出する。整形後のメールデータやコンテンツからのコンテキスト抽出については 4.2 節に示す。

④ メールからキー情報を取り出し、キー情報をもとにして、準備段階で導出した関数を参照する。

⑤ 参照した関数にメールコンテキストを入力し、入力から推測されるコンテンツコンテキストを得る。

⑥ 推測されるコンテンツコンテキストと、判定対象のメールに付されていたコンテンツとの類似度を算出する。

⑦ 類似度を閾値と比較し、不審メールであるかを判定する。

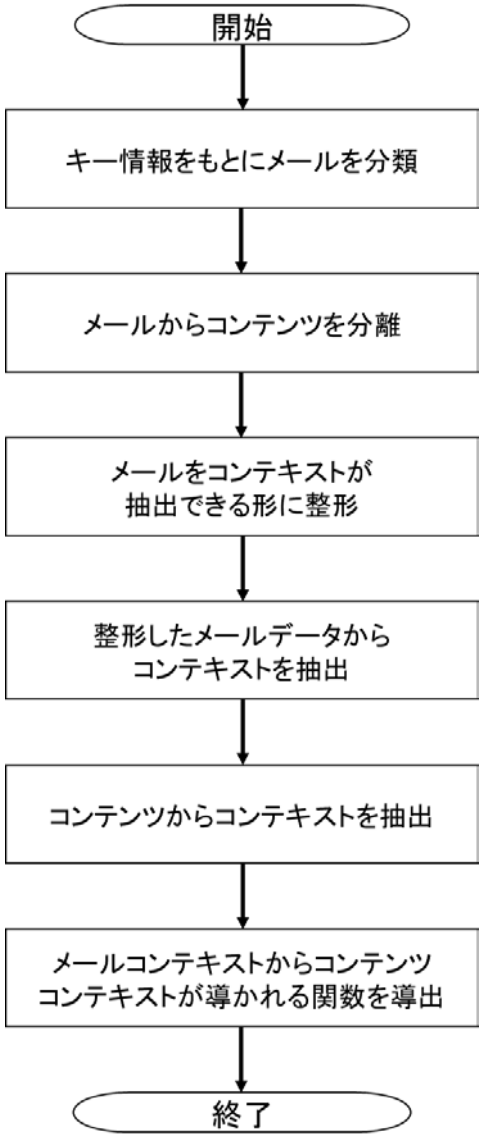

図 3 準備段階のフローチャート

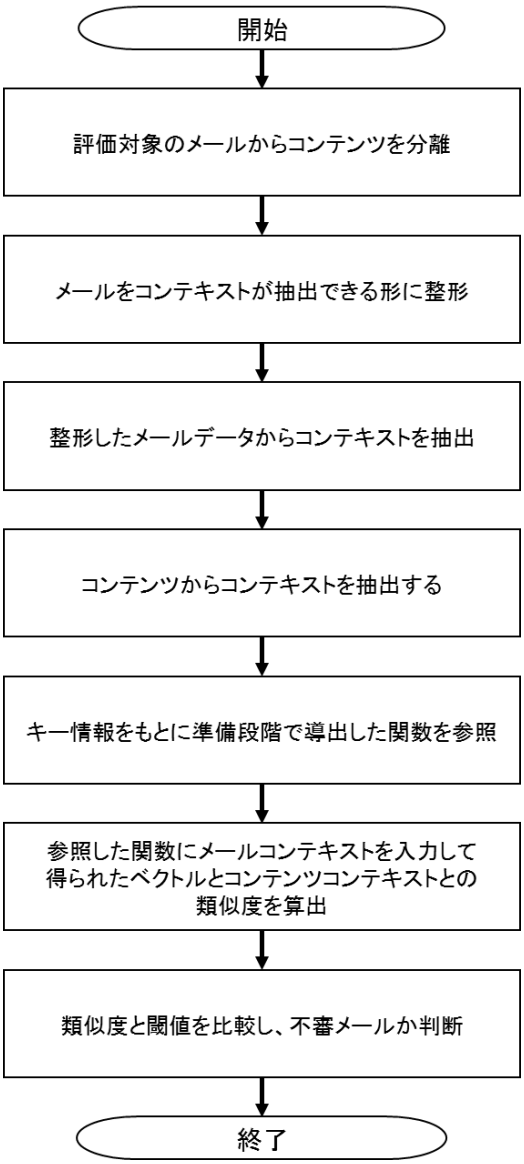

図 4 運用段階のフローチャート

4.2 コンテキストの抽出

本節では、メールのコンテキストと、コンテンツのコンテキスト、それぞれの抽出方法について示す。

メールのコンテキストは、メールから抽出可能な特徴ベクトルの連結によって表現される。整形後のメールデータである題名、宛先情報、本文、の三つの要素を特徴ベクトルに置き換えたのち、それらを連結したものをコンテキストとする。

各要素から特徴ベクトルを抽出する方法を、宛先情報と、題名や本文のような文章のそれぞれで示す。

宛先情報の特徴ベクトルへの変換は、現在抽出しようとしているメールのキー情報に含まれる宛先の一つ一つを、宛先情報が含むか含まないかでベクトル化する。

例えば、キー情報が「xxx@ab.com」、「yyy@ab.com」、「zzz@ab.com」、「abc@xx.com」の四つで、宛先情報が「xxx@ab.com」、「zzz@ab.com」、「efg@xy.com」の三つだとする。この場合、宛先情報の特徴ベクトルは、式(1)のようになる。

$$\vec{v} = (1,0,1,0) \tag{1}$$

題名と本文のような文章は、文章をベクトルに変換することができる doc2vec のような自然言語処理技術を用いることで特徴ベクトルを得る[5]。

他にも、TF-IDF のようなキーワード抽出技術により抽出したキーワードを BoW によりベクトル化してもよい。

以上の手続きにより、メールから得られる特徴ベクトルを式(2)に示す。

$$\vec{v} = \vec{a} \bullet \vec{b} \bullet \vec{c} \tag{2}$$

ただし、演算子●は、ベクトルの要素を結合する演算子であり、$\vec{a}$ は宛先情報の特徴ベクトル、$\vec{b}$ は題名の特徴ベクトル、$\vec{c}$ は本文の特徴ベクトルを示す。

コンテンツのコンテキストを抽出する方法は、コンテンツごとに異なる。例えば、コンテンツが PDF 形式の文書

ファイルであった場合、PDFMiner[6]のようなツールを用いることで、PDF に記載されている文章やファイル名を抽出することが可能である。文章を抽出後は、メールの題名や本文と同様に、doc2vec のような自然言語処理技術を用いることで特徴ベクトルを得る。

4.3 コンテキスト間の関係性学習

本節では、コンテキスト間の関係性学習について示す。

メールコンテキストの集合を C_m とし、その要素を c_{mi} とする。さらに、当該メールに対応するコンテンツコンテキストの集合 C_c とし、その要素を c_{ci} とする。式で表すと、次の式(3)、(4)、(5)、(6)のようになる。

$$c_{mi} \in C_m \qquad (0 \le i \le N) \tag{3}$$

$$c_{ci} \in C_c \qquad (0 \le i \le N) \tag{4}$$

$$c_{mi} = (x_{i1}, x_{i2}, \cdots, x_{iL}) \tag{5}$$

$$c_{ci} = (t_{i1}, t_{i2}, \cdots, t_{iM}) \tag{6}$$

ただし、N はメール集合の要素数で、c_{mi} は L 次元のベクトルとし、c_{ci} は M 次元のベクトルとする。

続いて、c_{mi} を入力として、最終的に c_{ci} が導かれるように学習する関数 f の要素を式(7)に示す。

$$f(c_{mi}) = c_{yi} = (y_{i1}, y_{i2}, \cdots, y_{iM}) \tag{7}$$

関数 f を確率的勾配降下法により学習させるための損失関数 E の例を、式(8)に示す。

$$E(c_{ci}, c_{yi}) = -\frac{1}{B}\sum_i \sum_k t_{ik} \log y_{ik} \tag{8}$$

ただし、B はメール集合の内から学習で用いるために選択したバッチ数である。

以上の式をもとに学習させた関数 f を、メールコンテキストとコンテンツコンテキストの関係性として、コンテンツ関係性データベースに登録する。

4.4 コンテキストの関係性比較

本節では、受信したメールのコンテキスト c'_m をもとに、受信したメールのコンテンツのコンテキスト c'_c との類似度比較について示す。

受信したメールから抽出したキー情報をもとに、準備段階で導出した関数 f を参照する。参照した関数 f に対して、受信したメールのコンテキスト c'_m を代入し、メールコンテキストの f による写像 c'_y を得る。c'_m から、c'_y を得る式を(9)に示す。

$$f(c'_m) = c'_y = (y'_1, y'_2, \cdots, y'_M) \tag{9}$$

続いて、得られた c'_y と、c'_c とを、二つのベクトルの類似度を評価する評価関数 g に代入し、閾値 th と比較して、類似しているかどうかを判定する。評価関数 g は、例えばコサイン類似度を用いる。コサイン類似度を用いた評価関数 g を式(10)に示す。

$$g(c'_c, c'_y) = \frac{c'_c \cdot c'_y}{|c'_c||c'_y|} \tag{10}$$

th と比較して、類似度が低い場合、評価対象のメールに付されたコンテンツは、メールの内容にそぐわないコンテンツであると判断できるため、不審なメールであると判断する。

5. 考察

5.1 学習データが不足する場合について

本手法で検知するためには、キー情報により分類される集合の、コンテンツ付のメールが一定数必要となる。これまでのやり取りが少ないグループや、新たに始まったプロジェクトでは、蓄積されている情報が少ないため、本手法を適用することが困難であると考えられる。解決策としては、蓄積情報が足りない場合には、送信者が送るファイルの特徴や、メールの内容から推測されるファイルの種類が、議事録や説明資料等であるかを推測する等により、提案手法を補うことが考えられる。

5.2 やり取りがある場合の誤検知について

今回、コンテンツが付されたメールだけに着目し、メールが不審であるかを判定した。しかし、添付ファイル等を送る際には、これまでのやり取りから類推できる代名詞や、やり取りをしている者同士であれば推測できる用語を本文に用いる場合もある。その場合には、仮に正規のメール送信であっても、不審であるとし、誤検知が発生する可能性が高い。解決策としては、診断対象に関しては、やり取りを引出し、代名詞等を補完した後に、提案手法を適用することが考えられる。

6. おわりに

本書では、メールのコンテキストと、メールに付されたコンテンツのコンテキストとに相関があることに着目し、正常なメールのコンテキストと、メールに付されたコンテンツとの関係性を学習することで、メールの文面は似せられているが、送信されているコンテンツに相関が無いようなメールを検知する手法を提案した。今後は、メールの学習データを揃え、本手法の有効性を検証する。

謝辞 本研究を進めるにあたり、アドバイスいただいた西垣先生と、西垣研究室に感謝いたします。

参考文献

[1] トレンドマイクロ：COMBATING MALICIOUS EMAIL AND SOCIAL ENGINEERING ATTACK METHODS, https://www.trendmicro.com/cloud-content/us/pdfs/business/datasheets/ds_social-engineering-attack-protection.pdf.
[2] CipherCraft/Mail, https://www.ntt-tx.co.jp/products/ccraftmailtypeh/
[3] Disarm, https://support.symantec.com/en_US/article.HOWTO93096.html
[4] Sevtap Duman, Kubra Kalkan Cakmakciy, Manuel Egelez, William Robertson and Engin Kirda, “EmailProfiler: Spearphishing Filtering with Header and Stylometric Features of Emails”, Computer Software and Applications Conference (COMPSAC), 2016 IEEE 40th Annual.
[5] Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean., “Distributed Representations of Words and Phrases and their Compositionality”, In Proceedings of NIPS 2013.
[6] PDFMiner, http://www.unixuser.org/~euske/python/pdfminer