「マルチメディア，分散，協調とモバイル
(DICOMO2018)シンポジウム」 平成30年7月

攻撃者のメール送信状態推定による不審メール検知技術の提案

西川弘毅[1,2] 山本匠[1] 河内清人[1] 西垣正勝[2]

概要：企業の担当者等と複数回のやり取りを通して信頼を得て，その後に添付ファイルのクリックや指定口座への送金を行わせるような攻撃が存在する．特に，昨今では企業の決済担当者に対して，その企業の重役や，取引相手に成りすまし，攻撃者の口座へ金銭を振り込ませるビジネスメール詐欺(BEC: Business E-mail Compromise)が，重大な脅威となっている．しかし，既存の技術では，巧妙な攻撃者によるやり取りを通じた攻撃メールを検知することができない．そこで本稿では，メールを受信した際に，そのメール送信者とのやり取りを抽出し，攻撃が行われる状態が来た場合に不審なメールであると検知することで，巧妙な攻撃の検知を行う技術を提案する．

Proposal of suspicious e-mail detection with estimated attacker state

HIROKI NISHIKAWA[1,2] TAKUMI YAMAMOTO[1]
KIYOTO KAWAUCHI[1] MASAKATSU NISHIGAKI[2]

1. はじめに

企業の担当者等と複数回のやり取りを行い，信頼を得た後に，添付ファイルをクリックさせることや，指定口座への送金といった目的を達成する種類の攻撃があり，重大な脅威となっている[1]．特に，企業の決済担当者に対して，その企業の重役や，取引相手に成りすまし，攻撃者の口座へ金銭を振り込ませるビジネスメール詐欺(BEC: Business E-mail Compromise)は，被害額が多くなる傾向があり，重大な脅威となっている[2][3].

また，報告件数は少ないものの，やり取りの後に標的がメールの添付ファイルの開封や，URL をクリックするように仕向ける，やり取り型攻撃も存在する．このような標的型攻撃は，依然脅威として存在する[4].

既存の対策として，不審メールを検知する手法や，ビジネスメール詐欺を防ぐための対策が存在するが，巧妙な攻撃者によって，話術巧みに回避される危険がある．他に，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[5]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[6]や，文章の個人識別技術を応用することで，届いたメールが，確かに送信者から送られたものであるかを個人識別することで，成りすましメールを検知する技術[7]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．例えば，企業の問い合わせ窓口に対して商品の問い合わせを行い，正常なやり取りを行う人物であると信頼を獲得したのち，攻撃者の目的であるマルウェア感染等を引き起こすメールを送るような攻撃を考える．このような攻撃の場合，攻撃者が外部のアドレスからメールを送信することはありうるため，送信者ドメイン認証や送信経路から不審であると判断することはできない．さらに，本人性を検証しても，初めから攻撃者本人からのメールであるため，本人であるかどうかでは攻撃を判断することができない.

そこで本稿では，メールを受信した際に，そのメール送信者とのやり取りを抽出し，受信したメールが，やり取りを見るに，被害が発生するような攻撃のメールである可能性がある場合に，不審なメールであると検知することで，巧妙な攻撃の検知を行う技術を提案する．

本稿では，2 章で関連研究について示し，3 章では，今回の手法による検知対象である，ビジネスメール詐欺と，やり取り型攻撃について説明する．4 章では提案する検知手法について説明する．5 章で，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，既存の不審メール検知技術について説明する．CipherCraft/Mail[5]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.

1 三菱電機株式会社 情報技術総合研究所
2 静岡大学 創造科学技術大学院

Disarm[6]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[7]は，個人ごとに，メール文面に特徴が存在することを利用し，不審なメールを検知する手法を提案している．本手法では，まず不審であるかの識別対象である個人ごとにメールを収集し，個人ごとの特徴量を SVM で学習する．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ことで，届いたメールが，正しく本人からの文章であるかを判断し，不審な成りすましメールを検知することができる．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，という課題がある．

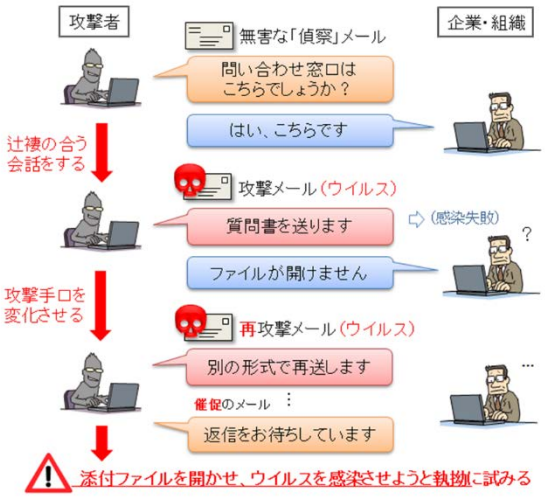

図 1　やり取り型攻撃の例

3. 本方式による検知対象

本章では，本稿でスコープとする攻撃である，ビジネスメール詐欺と，やり取り型攻撃の概要を説明する．

3.1 ビジネスメール詐欺

ビジネスメール詐欺(BEC)は，巧妙な騙しの手口を利用して，メールによって金銭をだまし取る攻撃である．具体的には，攻撃者が，攻撃対象とする企業の重役や取引関係のある企業の担当者になりすまし，攻撃対象企業の決済担当者にメールを送り，攻撃者の口座へ送金するように誘導する．

巧妙なビジネスメール詐欺においては，技術的な対策が難しく，一人ひとりが手口を理解し，“怪しさ”を見抜くことが重要である，と言われている[2].

3.2 やり取り型攻撃

本節では，IPA の資料[4]を参考にし，攻撃者が標的とやり取りを行うことで信頼を得た後，感染行動に移るやり取り型攻撃の説明を行う．

資料[4]中に，「「やり取り型攻撃」とは，一般の問い合わせ等を装った無害な「偵察」メールの後，ウイルス付きのメールが送られてくるという，標的型サイバー攻撃の手口の一つです.」と，記載されている．攻撃者は，対象とする組織の外部向け窓口等に対して，返信せざるを得ないメールを送りつける．対象から返信があると，辻褄の合う会話をしながら，マルウェアである添付ファイルを開かせ，組織へのマルウェア感染を試みる．やり取り型攻撃のイメージを図 1 に示す．

4. 提案手法

本手法によって検知する対象は，受信したメールのヘッダ情報や，添付ファイルから不審さを検知できない上に，やり取りを行う，やり取り型攻撃のメールである．ヘッダ情報や，添付ファイルの不審さから，不審であると判断できる攻撃メールは，関連研究等の手法により検知できているものとする．

本手法では，メールを受信した際に，そのメール送信者とのやり取りを抽出し，攻撃が行われる状態が来た場合に不審なメールであると検知することで，巧妙な攻撃の検知を行う．

本手法で実施する処理の全体像を以下に示す．また，図 2 に全体像を図示する．

本手法は大きく，検知ルールの設定を行う準備フェーズと，実際に不審メールの検知を行う運用フェーズの二段階により構成される．

(1) **準備フェーズ: 作業者により，検知に用いるメールやり取りのルールを登録する.**

(2) **運用フェーズ: メールを受信した際に，これまでのメールのやり取りを考慮し，受信したメールが不審であるかを判断する.**

以降の節では，各フェーズにおける処理を説明する．

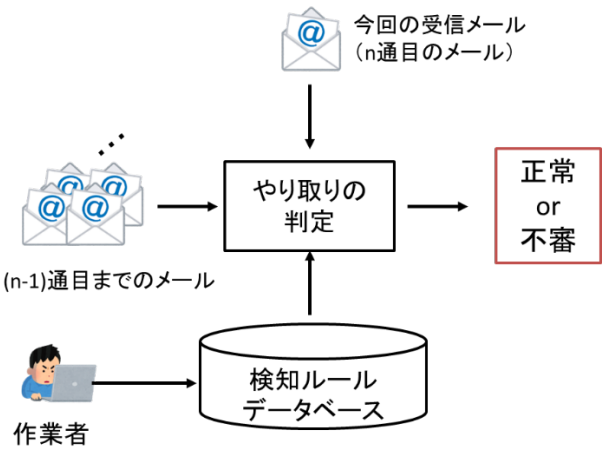

図 2　提案手法の全体像

表 1　状態の定義例

状態	ラベル	特徴量
s_1	窓口確認	(0.3,0.5,……,0.9)
s_2	承認	(0.1,0.8,……,0.7)
s_3	添付ファイル開封	(0.6,0.3,……,0.5)

4.1 準備フェーズ

準備フェーズでは，次の二つの処理を実施する．

(1) 状態の定義：メールの文章から得られる特徴量とラベルにより，状態を定義する．

(2) 検知ルールの登録：定義した状態を用いて，検知ルールを作成する．

以降の項では，各処理について説明する．

4.1.1 状態の定義

検知ルールを記述する際に用いる状態を定義する．状態は，それぞれ，状態が何を示しているかのラベルと，その特徴量とのペアにより定義される．それぞれの状態は，1 行目の状態であれば，s_1のように表現する．状態定義の際には，ラベルを示すような一般的なメール文章を選別し，その文章を Word2Vec[8]や Sentence2Vec[9]のような，文字から分散表現を獲得する技術を用いることで，特徴量を得る．状態の定義例を表 1 に示す．

4.1.2 検知ルールの登録

状態を定義した後に，メールが不審であるかを判定する処理を行う準備として，検知ルールを登録する．検知ルールの登録処理は，予め検知のためのルール集合$\mathcal{R}$を登録することである．ルール集合$\mathcal{R}$は，各ルールr_iを構成要素とする集合である．ルールr_iは，状態s_jを単語とした正規表現により記述する．ルールを正規表現で記述できることで，例えば似た文面を何度か繰り返した後に攻撃行動を行うようなやり取りのルールや，信頼を構築するためのあて先確認のメールの後，更に催促をすることがある場合も，無い場合も両方とも存在する際にも，簡単に表現することができる．ルールは，攻撃のやり取りとして想定されるメールスレッドをもとに，手作業で作成する．検知ルール登録の流れを図 3 に示す．ここで，図中のルール r_1 ，r_2は正規表現により記載されており，次のような意味を持つ．

r_1：０～１回の s_1の後，１回の s_3

r_2：０～２回の s_1の後，　０～１回の s_2の後，　１回の s_3

ルールを作成する際には，例えば既存のビジネスメール詐欺で用いられる手法[2]や，ビジネスにおいて，取引が行われる際には必ず実施される手続きをモデル化し，どのタイミングであれば攻撃が可能であるか，というサイバーキルチェーンと似た形にできればよいと考えている．

4.2 運用フェーズ

運用フェーズでは，診断対象のメールを受信した後，次の処理を実施する．運用フェーズの全体像を図 4 に示す．

(1) メールスレッドの取得：受信したメールと，受信したメールと関連するやり取りのメールから，メールスレッドを構築する．

(2) スレッドを状態系列に変換：メールスレッドの各メールを，状態へと変換することで，メールの代わりに状態が並んだ，状態系列を得る．

(3) 状態系列とルールの比較：ルール集合の全要素と状態系列を正規表現により比較し，一致した場合は不審と判断する．

以降の項では，各処理について説明する．

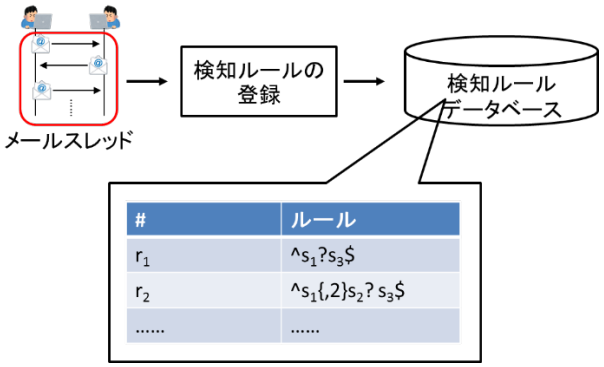

図 3 検知ルール登録の流れ

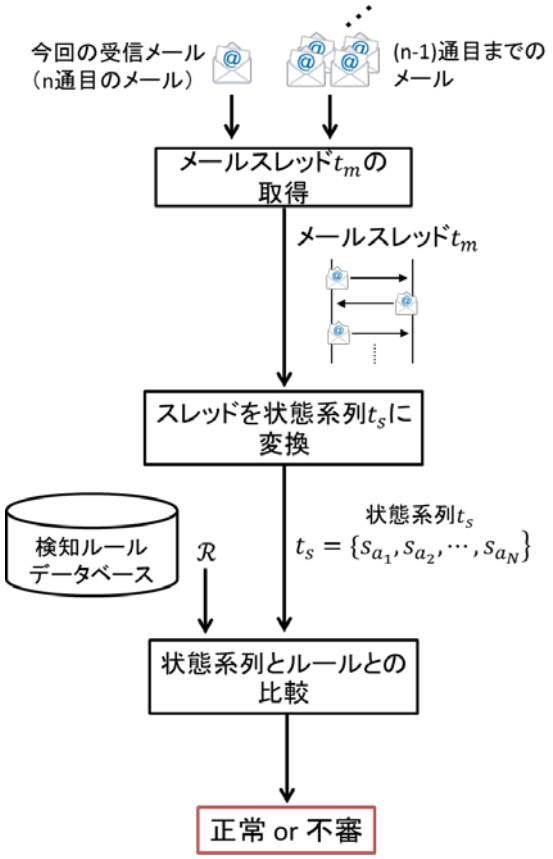

図 4　運用フェーズの全体像

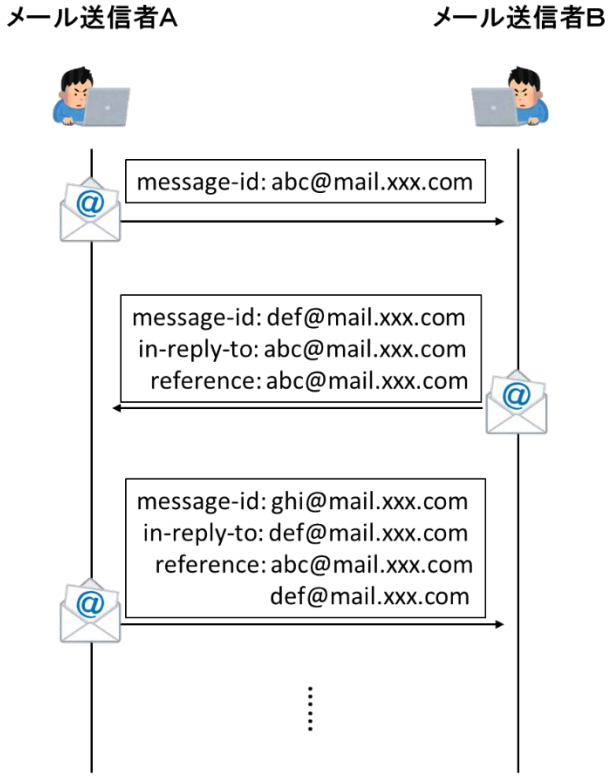

図 5　メールやり取り時にメールヘッダに付帯する情報

4.2.1 メールのスレッド化

メールを，やり取りの順番に並べることで，スレッドかする．メールのスレッド化に必要な情報については，RFC2822 に記載がある[10]．メールスレッドの構築は，例えば Message-Id と，In-Reply-To，References というヘッダ情報を参照することで構築することが可能である．各メールには Message-Id が付与されている．メール返信時に，In-Reply-To と References の値が，返信対象である Message-Id の値で更新されるため，上記ヘッダを参照することで，スレッドを構築することが可能となる．

4.2.2 スレッドの状態系列への変換

本処理では，メールスレッドt_mから，準備フェーズで定義した状態が並んだ列である，状態系列t_sへと変換する．以下の順に処理を実施する．

1. 得られたメールスレッドt_mの各メールm_iを特徴ベクトルv_iに変換する
2. 定義した状態s_jの中から，最もv_iと，s_jの特徴量，との距離が近い状態を選択する
3. メールm_iと状態s_jとを対応づけ，状態系列t_sを得る

4.2.3 状態系列とルールの比較

本処理では，状態系列t_sが，ルール集合の要素と一致するかを確認することで，メールが不審であるかの判断を行う．本処理の概要を図 6 に示す．

本処理では，正規表現のマッチングを行う関数である，$match(\mathrm{r,t})$を用いて，ルール集合$\mathcal{R}$の全要素に対してスレッドの状態系列による表現tを引数として比較処理を行う．ただし，第一引数rに正規表現のパターンを，第二引数tに比較対象の系列を代入するものとする．正規表現の比較処理で，一致したルールがあった場合，不審なメールと判断する．全てのルールが，状態系列と一致しない場合には，正常であると判断する．

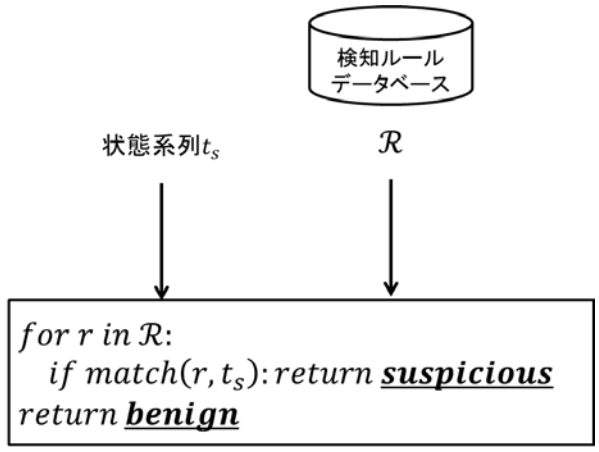

図 6　状態系列とルールとの比較

5. 考察

5.1 未知のやり取りについて

テンプレートとして登録していない未知のやり取りを検知することはできない.

そこで，新しく発生した攻撃は，テンプレートとして公開し,共有することで,攻撃を防ぐことができると考える．現在でも，新たな手口による攻撃メールが行われた際に，文面ややり取りのケースを例示しているが，これはメールを受け取った人物が解釈する必要があり，正確に展開することも難しい．本方式によって，ルールを共有することができれば,やり取りによる攻撃の情報が共有された際にも，やり取りが不審であるかを，セキュリティのレベルが高い人間や，疲れているとき等で正常な判断が行えない場合であっても，不審なメールを検知することができるようになる．テンプレートを公開する際に注意する点としては，誤検知につながらないように注意することである．

5.2 誤検知への対応

提案手法では，正規のメールであってもルールに合致する場合には不審なメールであると判断してしまうため，誤検知の可能性がある．そのため，ルール登録の際に，過去の攻撃手法と関連付けて登録することで，不正であると検知した際に，どの攻撃と類似しているかをシステムで提示できるようにすることを考えている．この場合，誤検知は減らないため，誤検知が多く発生してしまう場合はこの対策は有効ではないが，誤検知の数が限られる場合には，ユーザに気づきを与えることで攻撃にかかる確率を下げることができると考えている．

5.3 メールの特徴量変換について

提案手法では，メールがルールに一致するかどうかを文章の分散表現による特徴量変換を用いている．しかし，分散表現への変換は，その変換器の学習データに依存し，更に企業名等のキーワード等の一般的ではない表現は特徴量をうまく算出することができない．そのため，分散表現を行う前に，メール本文に対して，企業名等はリプレース処理を実施し，活用が可能な表現は，活用形を終止形に統一する等の事前処理を実施することで，特徴を抽出しやすくすることが考えられる．

6. おわりに

本稿では，巧妙な攻撃の中でもやり取り型攻撃を検知するための手法として，やり取りの状態をテンプレートとすることで，巧妙な攻撃を検知する手法を提案した．今後は実際に攻撃メールが検知可能であると考えられるルールを作成し，本方式を実装した後，評価を行う．

参考文献

[1] IPA, 情報セキュリティ 10 大脅威 2018, https://www.ipa.go.jp/security/vuln/10threats2018.html

[2] IPA, 【注意喚起】偽口座への送金を促す“ビジネスメール詐欺”の手口, https://www.ipa.go.jp/security/announce/20170403-bec.html

[3] YOMIURI ONLINE, ＪＡＬ３．８億円詐欺被害　ビジネスメールに割り込み偽請求, http://www.yomiuri.co.jp/science/goshinjyutsu/20180109-OYT8T50178.html

[4] IPA, 組織外部向け窓口部門の方へ :「やり取り型」攻撃に対する注意喚起 ～ 国内 5 組織で再び攻撃を確認 ～, https://www.ipa.go.jp/security/topics/alert20141121.html.

[5] CipherCraft/Mail, https://www.ntt-tx.co.jp/products/ccraftmailtypeh/

[6] Disarm, https://support.symantec.com/en_US/article.HOWTO93096.html

[7] 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.

[8] Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean., “Distributed Representations of Words and Phrases and their Compositionality”, In Proceedings of NIPS 2013.

[9] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International Conference on Machine Learning. 2014.

RFC2822, “”http://www.emaillab.org/emailref/RFC/rfc2822.txt