社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－DPS－128 (4)
2006－GN－ 61 (4)
2006－EIP－ 33 (4)
2006／9／14

Webサイトからの企業名抽出による フィッシング対策手法の提案

柴田 賢介† 荒金 陽助† 塩野入 理† 金井 敦†
† 日本電信電話株式会社 NTT 情報流通プラットフォーム研究所

概要 電子メール中のハイパーリンクから偽装サイトに誘導し，個人情報を詐取するフィッシングの被害は増加する傾向にあり，EC だけでなく，社会／行政サービスの電子化への影響も懸念される．昨今では，フィッシングの誘導手段は電子メールに限らず，ブログのトラックバックや，IM(Instant Messenger) 等に偽装サイトへの URL を貼り付けるケースも散見される．本論文では，Web ページが騙る企業名を抽出し，当該企業名に対応する正当な URL のリスト (ホワイトリスト) を用いて個人情報の送信先となる URL の正当性を検証するフィッシング対策手法を提案する．また，本手法実現の鍵となる，Web ページからの企業名抽出アルゴリズムについて説明する．

キーワード EC，社会／行政サービス，情報セキュリティ，フィッシング，ソーシャルエンジニアリング

An Anti-Phishing Method by Detecting Brand Names in Websites

Kensuke Shibata† Yosuke Aragane† Osamu Shionoiri† Atsushi Kanai†
† NTT Information Sharing Platform Laboratories, NTT Coporation

Abstract Cybercrimes which aim at users' personal information become a serious threat. One of these crimes is a phishing attack. Phishing attack may have negative effects on the spread of e-commerce or e-governance. In recent years, the lure of the phishing skips from emails to other media(ex. Instant Messenger, hyperlink on attachments to email messages). In this paper, we propose an anti-phishing method. This proposal method has a function of hijacked brand name detection from websites. This method also has the function of URL verification by using white list of legitimate web sites. We show the architecture of our method and algorithm of hijacked brand name detection.

Keywords Electronic Commerce, Social/Administrative Services, Information Security, Phishing Attack, Social Engineering

1 はじめに

近年のインターネットの普及により，オンラインショッピングやオンラインバンキングなどの電子商取引や，電子投票，電子申請といった電子政府への取り組みが拡大しつつある．これらのサービスの普及により，利用者に対する利便性が高まる一方で，インターネット上でやりとりされる個人情報を狙う犯罪が多発している．このような犯罪の１つに，フィッシングがある．フィッシングとは，図１に示すように，詐称メール (フィッシングメール) によってエンドユーザを偽装 Web サイト (フィッシングサイト) へ「誘導」し，フィッシングサイトにおいてクレジットカード番号や ID，パスワードを入力，送信させて個人情報を「詐取」するというものである．

フィッシング対策について取り組んでいる米国の団体 Anti-Phishing Working Group (APWG) によると，2006 年 5 月の 1ヶ月間に報告された新たなフィッシングサイトの件数は 11,976 件となっており，増加の一途をたどっている [1]．また，フィッシングサイトの平均寿命は 5.0 日と言われている．第三者の Web サイトを乗っ取るなどしてフィッシングサイトを開設し，個人情報を短期間で詐取した後，サイトを削除することにより，証拠をほとんど残すことなく，個人情報を詐取することが可能となっている．

本研究では，フィッシングサイトにおける個人情報の詐取を阻止することを目的とし，Web サイトが騙る企業名を抽出し，当該企業名に対応する正当な URL のリスト (ホワイトリスト) を用いて個人情報の送信先となる URL の正当性を検証するフィッシング対策手法を提案する．本論文ではまず，２章において関連研究について述べ，３章において，企業名を抽出した上でホワイトリストによる URL の正当性検証を行なうことの効果と，提案手法を実現するシステムのアー

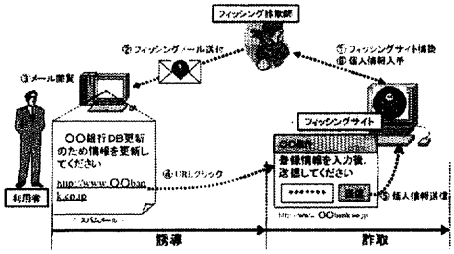

図 1: フィッシングの概要

キテクチャについて説明する．4 章では，Web サイトから企業名を抽出するアルゴリズムについて説明し，最後に，5 章にて本論文をまとめる．

2 関連研究

現在，フィッシングの手口として，エンドユーザにフィッシングメールやフィッシングサイトが正当なものであると信じ込ませるためのソーシャルエンジニアリングがより洗練される傾向にある．ソーシャルエンジニアリングにより，フィッシング詐欺師はエンドユーザとの間に顧客と企業という人間関係を巧みに作り出し，エンドユーザがフィッシングサイトにおいて自ら個人情報を入力するように仕向けている．Dhamija らは，(1) エンドユーザの知識不足の悪用，(2) 視覚的な偽装，(3) エンドユーザの不注意を悪用，といった 3 種類の手法を用いたフィッシングサイトを作成し，被験者に対して真贋判定を行なわせる実験を実施している [2][3]．実験結果によると，被験者の約半数が，Web サイトのロゴやレイアウト，デザインといったサイトのコンテンツのみから真贋を判断しているか，サイトのコンテンツとアドレスバーのドメインのみから判断しているという結果が得られている．アドレスバーに表記される文字列については，本物と 1 文字違いのドメイン名を取得したり，Javascript を用いてアドレスバーを偽装するといった手法によってエンドユーザを騙すことが可能である．企業の Web サイトは，当然のことながら全世界に発信されているものであり，HTML のソースファイル，企業のロゴ画像等は容易に入手することが可能である．つまり，これらをコピーすることにより，本物と酷似したフィッシングサイトを簡単に作成することができる．Dhamija らの実験結果を鑑みると，今後ソーシャルエンジニアリングがより洗練されることにより，フィッシングの被害は拡大することが推測される．

柴田らは，ソーシャルエンジニアリングへの対処を目的としたフィッシング対策手法に関する提案を行なっている [4]．ソーシャルエンジニアリングの手口においては，フィッシングメールやフィッシングサイトを正当なものであるとエンドユーザに信じ込ませる必要があることから，メール中にフィッシングのターゲットとなる企業名が出現する傾向がある．筆者らはこれを逆に利用した対策を提案している．図 1 における，フィッシングメールからフィッシングサイトへの誘導を阻止するために，フィッシングメールが騙る企業名を抽出し，企業名を特定した上で，正当な URL のリスト (ホワイトリスト) とメール中のリンク先 URL とを比較し，リンク先 URL の正当性を検証する．検証によりリンク先 URL が正当でないと判断された場合には，リンク先 URL への遷移をブロックする．しかし，フィッシングサイトへの誘導手段はメールに限られておらず，IM(Instant Messenger) やブログのトラックバック，メールの添付ファイル中に記載されたハイパーリンクから誘導する手口等が現れているため，フィッシングメールからの誘導をブロックするだけではフィッシング対策として十分であるとは言えない．

3 提案するフィッシング対策

本研究では，図 1 におけるフィッシングサイトでの個人情報の詐取をブロックすることを目的とし，Web サイトにおいて個人情報を送信する時点で，送信先の URL を検証することによるフィッシング対策を提案する．

3.1 企業名抽出とホワイトリストによるフィッシング対策

1 章において述べたように，フィッシングサイトには寿命が短いという特徴がある．フィッシング対策技術の中には，フィッシングサイトを発見し，サイトの URL をブラックリスト化することによって，リスト中の URL にアクセスした場合にエンドユーザに対して警告を行なう手法がある．しかし，フィッシングサイトの寿命が短いために，フィッシングサイトの URL を発見し，ブラックリストに掲載された時にはすでにフィッシングサイトが削除されていた，といったケースも考えられる．ブラックリスト方式は，リストの有効性を保つために，世界中のフィッシングサイトを見つけ出し，サイトの URL をリアルタイムにリスト化

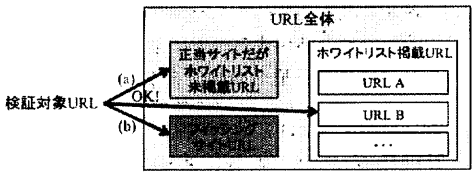

図 2: ホワイトリストにおける課題

するという処理を継続的に行なう必要があり，コスト面で課題があると言える．

本研究では，ブラックリストとは逆の概念である，正当な URL のリスト (ホワイトリスト) によるフィッシング対策を提案している．ホワイトリストは，Web サイトの構成が変更された場合にのみリストを更新すれば良いため，ブラックリストに比べると更新の頻度は低い．しかし，単純に正当な URL を羅列したホワイトリストには課題がある．ホワイトリスト中にリンク先 URL が存在しなかった場合，リンク先 URL は「正当であるとは言えない」という判断にとどまり，「フィッシングサイトである」とは言えない．これは，リンク先 URL がホワイトリストに未掲載の正しい URL なのか (図 2 の (a))，あるいはフィッシングサイトの URL なのか (図 2 の (b)) を判別することが不可能だからである．システムが「正当であるとは言えない」という判断を下してしまうと，最終的な URL の正当性判断をエンドユーザが行なうことになる．Web サイトの正当性に関する判断をエンドユーザに行なわせることの危険性は文献 [2] において述べられているとおりである．

本手法では，Web サイトを騙る企業名を特定した上で当該企業名に紐付けられたホワイトリストを用いて送信先 URL の正当性検証を行なう．これにより，検証に失敗した場合には，「検証対象の URL は当企業の正当な URL ではなく，フィッシングサイトである」と判断することが可能であり，エンドユーザの判断を必要とせず，システムにおいて「リンク先 URL はフィッシングサイトである」と言うことができる．

3.2 システムアーキテクチャ

本節では，提案手法を実現するシステムのアーキテクチャと動作について述べる．図 3 に，システムのブロック図を示す．エンドユーザ端末側のソフトウェアは，Web サイトにおける個人情報の送信を契機として動作し，企業名の抽出，送信先 URL の検証を行なう．

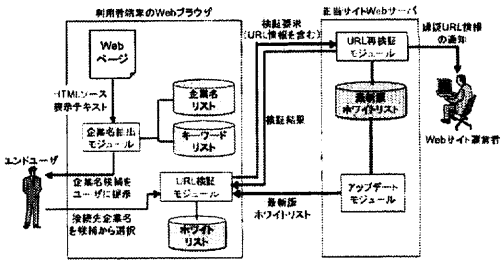

図 3: 提案手法を実現するシステムのブロック図

う．エンドユーザ端末における URL 検証に失敗した場合には，正当サイトの Web サーバに対して問い合わせを行ない，最新版ホワイトリストによる URL 検証を行なうとともに，ホワイトリストの更新が実行される．

以下に，エンドユーザが Web ページ中の送信ボタンを押すなどして，個人情報を送信しようとした場面を想定し，処理手順を示す．

1. Web ブラウザにおいて Web ページを読み込んだ際に，HTML のソースと，ブラウザ上に表示されているテキスト (以後，本テキストを表示テキストと呼ぶ) を取得する．

2. HTML ソース，表示テキストに対して企業名抽出アルゴリズム (4 章を参照) を適用し，各企業名がエンドユーザの意図する接続先の企業名となる度合いに応じて得点を付与する．(以後，エンドユーザが意図する接続先の企業名を接続先企業名と呼ぶ)

3. 得点が閾値を超えた企業名を接続先企業名の候補としてエンドユーザに提示し，接続先企業名の選択を求める．候補が複数存在する場合には得点順に提示する．

4. エンドユーザが選択した接続先企業名に対応するホワイトリストに，送信先 URL が含まれるか否かを検証し，含まれない場合には接続先企業のサーバに対し，送信先 URL の情報を送信する．

5. 当該企業のサーバにおいて，再度ホワイトリストと送信先 URL との検証を行ない，ホワイトリストに送信先 URL が含まれない場合にはフィッシングサイトであると判断する．サーバ側で運用者に通知するとともに，エンドユーザに対しては検証結果を通知し，最新版のホワイトリストを送信する．

6. ホワイトリストの検証により，送信先 URL がフィッシングサイトであることをエンドユーザに通知するとともに，個人情報の送信をブロックする．

以上の手順により，エンドユーザが個人情報を送信しようとしている URL の正当性を確認した上で送信の可否を決定することが可能であり，エンドユーザは安全に Web ページでの個人情報送信を行なうことが可能となる．

4 企業名抽出アルゴリズム

3.2 節において述べた提案手法では，接続先企業名を正しく抽出することによって，検証すべきホワイトリストが特定され，送信先 URL の検証が可能となる．つまり，接続先企業名の抽出精度が本システムの核となっていると言える．本章では，Web ページを騙る企業名を抽出するアルゴリズムについて説明する．

4.1 企業名抽出アルゴリズムの要件

企業名抽出アルゴリズムは，エンドユーザがアクセスしている Web ページに対して適用され，抽出した企業名は接続先企業の候補としてエンドユーザに提示される．エンドユーザが行なう Web アクセスに対して安全性が保たれ，かつ接続先企業の選択時にエンドユーザに対して過度な負荷をかけない，という観点から，企業名抽出アルゴリズムには以下の三点が求められる．

1. 接続先企業名が漏れることなく候補として抽出され，エンドユーザに対して選択肢として提示されていること．

2. 利用者が選択を行なう際に多くの選択肢が表示されることによって混乱を招くことを防ぐため，確からしい候補に絞り込んだ上で選択肢を構成すること．

3. 利用者が容易に接続先企業を選択することができるように，順位付けによって接続先企業名が先頭に表示されること．

また，従来のフィッシング詐欺に用いられてきた手法を鑑みると，Web ページからの企業名抽出に対して以下のような攻撃が想定される．このような攻撃への耐性があることも企業名抽出アルゴリズムの要件となる．

i) 企業名を表す文字列をページ中に使用せず，企業のロゴ画像を多用することにより，テキストマッチングによる企業名の検索を不可能にする．

ii) Javascript を用いてエンコードした文字列を HTML のソースとして記述することにより，HTML のソース中に現れる企業名を隠蔽する．

iii) 背景と同系のフォントカラーを用いた文字や，極めて小さいフォントサイズの文字を使って Web サイトの内容と無関係の文字列を利用者に見えにくい状態でサイト内に含めることにより，テキストマッチングによる企業名の検索を攪乱する．

iv) 企業名を表す文字列の文字と文字の間に空白文字を挿入することにより，文字列単位での企業名の検索を不可能にする．

v) 利用者に個人情報を入力させるフォームを含むフィッシングサイトを正当な企業の Web サイト上にポップアップとして表示し，ポップアップ画面自体には企業名を含まない．

4.2 企業名抽出のためのルール

4.1 節において企業名抽出アルゴリズムの要件を述べた．これらの要件を満たすために，本手法では企業名抽出に関する複数のルールを定義し，これらのルールに則って Web ページから企業名を抽出する．

A) 企業の呼称をリスト化したもの (企業名リスト) を事前に用意しておき，Web ページの HTML ソースと当該リストに含まれる企業名とを比較することにより，企業名の出現回数をカウントする．出現回数が多い企業名に高い得点を付与する．4.1 節の i) において述べた，画像を多用するフィッシングサイトであっても，画像へのリンク URL 等に企業名が含まれることが多いため，企業名抽出の精度は高まると考えられる．

B) 上記の企業名リストに含まれる企業名と，Web ブラウザに表示されているテキスト (表示テキスト) を比較することにより，企業名の出現回数をカウントする．出現回数が多い企業名に高い得点を付与する．図 3 において Web ページから HTML ソースと表示テキストの 2 種類を取得しているのはこのためである．

C) Web ページの HTML ソースに対し，Javascript によってエンコードされている文字列をデコード

し，可読な形式に復元した後に A) の処理を行なう．これにより，4.1 節の ii) において示した攻撃手法への対策が可能となる．

D) Web ページの HTML ファイルに対し，HTML のタグ情報を考慮し，強調されていると推測される企業名に得点を付与する．考慮するタグ情報の例としては，以下が挙げられる．

- 他と異なるフォントカラーを使用している．
- 他より大きいフォントを使用している．
- 他と異なるフォントを使用している．
- title タグ等の強調を目的とするタグによって囲まれている．

逆に，以下の条件に該当する企業名については，得点を付与しない．これにより，4.1 節の iii) において示した攻撃手法への対策が可能となる．

- 背景色と同系のフォントカラーを使用している．
- 他より極めて小さいフォントを使用している．

E) Web ブラウザに表示されているテキスト情報の中の改行，スペース等の空白文字を削除した文字列に対し，B) の処理を行なう．これにより，4.1 節の iv) において示した攻撃手法への対策が可能となる．

F) Web ブラウザに表示されているテキスト情報の中に含まれる企業名が特定の位置にある場合に，得点を付与する．特定の位置とは，例えばテキスト情報の先頭，末尾部分，事前に用意されたキーワード（例：Copyright，編集，Subject，From）の前後等が挙げられる．

G) 過去に利用者が閲覧した Web ページにおいて既に抽出されている企業名を利用する．これにより，4.1 節の v) において，利用者が現在アクセスしている Web ページに企業名が含まれない場合に，過去の閲覧履歴数ページ分において抽出した企業名を利用することが可能となる．

H) 上記 A)〜G) のルールよって抽出された企業名に対し，ルール毎に重みを付与した上で，企業名毎に総得点を計算し，閾値と比較して利用者に提示する企業名を絞り込む．また，絞り込んだ企業名を総得点順にソートし，最も総得点の高い企業名をリストの先頭とする．

A)，B) については，企業名抽出の対象となる Web ページが正当である場合もしくはフィッシングサイトである場合の両者共に，HTML のソース中もしくは Web ブラウザに表示されているテキスト中に当該 Web ページを騙る企業名が数多く含まれることが多いため，これをカウントして得点を付与している．B) の Web ブラウザにおいて表示されているテキストを HTML ソースとは別に利用している理由は，4.1 節の ii) において示した Javascript によるエンコードに類する手口が現れた場合に，最終的に利用者が閲覧している Web ブラウザ上からのテキストを取得して企業名を抽出することにより，企業名抽出の精度向上が見込めるためである．

G) において事前に用意されるキーワードについては，多くの Web ページにおいて，著作権表示や文責を示す表記がページ中の下部に記述される傾向があるため，表記を行なう際に企業名とともに用いられる「Copyright」や「編集」等といったキーワードを含めておく．また，Web メールシステムでは，メールの題名や送信者は，例えば「Subject：○○」，「題名：○○」，「From：△△」，「送信者：△△」といった表記によって示される．メールの題名や送信者には当該メールを騙る企業名が含まれることが多いため，「Subject」，「題名」，「From」，「送信者」等のキーワードとともに用いられる企業名に注目することにより，利用者が意図する企業名の抽出が可能となる．

I) によって各企業名の総得点が算出され，閾値と総得点とを比較し，閾値を超えている企業名を絞り込み，さらに総得点の降順にソートした結果を利用者に提示することにより，4.1 節において述べた要件を満たす企業名抽出を行なうことが可能となる．

4.3 企業名抽出アルゴリズムの評価

本論文では，企業名抽出アルゴリズムによる接続先企業名の抽出精度に関する予備評価として，フィッシングサイトにおける企業名の出現回数を調査した．4.2 節の A) において，Web ページの HTML ファイル中に出現する企業名をテキストマッチングによりカウントするため，サイト中に接続先企業の名称が多く出現することにより，接続先企業を正しく抽出できる可能性が高くなる．

評価は 2006 年 5 月から 6 月の間に取得した実際のフィッシングサイトの HTML ソース 100 個を対象として実施した (但し，ハイパーリンクによる遷移が行なわれるフィッシングサイトにおいては，各々のペー

ジを1個としてカウントしている)．それぞれのサイトにおける接続先企業名をリストとして用意しておき，ソースと企業名リストとのマッチングにより，ソース中の企業名出現回数を調べた．評価の結果を図4に示す．フィッシングサイト1ページあたりの接続先企業名の出現回数は，平均45.9回となっており，1ページの中に接続先企業名が繰り返し使われていることが分かる．このように高い頻度で接続先企業名が現れた理由としては，以下の2点が挙げられる．

- ソーシャルエンジニアリングを成功させるために，Webページ内で企業名を多用し，正当な企業のWebサイトであると信じ込ませようとする傾向が強い
- ロゴ画像へのハイパーリンクのURLに企業名が含まれている．

後者については，正当な企業のWebサイトへ直接リンクを貼ることにより，ロゴ画像をフィッシングサイトに組み込んでいる場合や，フィッシングサイトを設置する対象となったサーバに企業名が含まれるディレクトリを作成し，ロゴファイルを格納している場合がある．企業名を含むディレクトリを作成してURL中に企業名を含めることによって，ソーシャルエンジニアリングの成功率を高くするという意図があると考えられる．

100個の評価対象ページのうち，接続先企業名が一度も現れないページが2個存在した．2個のサイトのうち，一方はescape関数を入れ子にして使用するなど，複雑な構造を持つサイトであり，HTMLソースをunescape関数でデコードした結果においても，接続先企業名は検出できなかった．つまり，4.2節のC)の手法を用いても，検出できないこととなる．しかし，Webブラウザ上に当該ページを表示すると，接続先企業名がテキストとして8箇所に現れており，4.2節のB)の手法を用いることにより，接続先企業名の取得が可能である．もう一方のページは，ページ中に企業名が全く含まれておらず，当該ページのみでの接続先企業名の取得は困難である．但し，当該ページはログイン画面から遷移して二番目に表示されるページとなっており，4.2節の手法G)によって，過去に閲覧したWebページから抽出した企業名を利用することにより，接続先企業名を抽出できると考えられる．

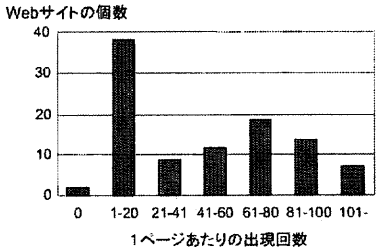

図4: 1ページあたりに出現する接続先企業名の数

5 まとめと今後の課題

本論文では，フィッシングサイトにおける個人情報の詐取をブロックすることを目的とし，Webページが騙る企業名を抽出した上で，当該企業名に対応するURLのホワイトリストを用い，個人情報の送信先となるURLを検証する手法を提案した．本手法は，企業名を特定した上でホワイトリストによるURL検証を行なうことにより，検証対象となるURLが正当なものであるのか，もしくはフィッシングサイトであるのかを明確に判断することが可能となっている．

本手法において，フィッシングサイト検知の成否は，Webページを騙る企業名を正しく特定できるか否かに依存している．そこで，現在のフィッシング手口を考慮し，Webページを騙る企業名を正しく抽出するための企業名抽出アルゴリズムを提案した．今後は，フィッシングサイトもしくは正当な企業サイトを対象とし，企業名抽出の精度に関する評価を行ない，抽出アルゴリズムにおいて使用している複数のルールへの重み付けの最適化を行なう．

参考文献

[1] Anti-Phishing Working Group: Phishing Activity Trends Report - May, 2006, http://www.anti-phishing.org/reports/apwg_report_May2006.pdf (2006).

[2] Dhamija, R., J.D.Tygar and Hearst, M.: Why Phishing Works, *Conference on Human Factors in Computing Systems(CHI2006)* (2006).

[3] 荒金陽助, 間形文彦, 柴田賢介, 塩野入理, 金井敦: フィッシング詐欺対策に関わる研究動向と法適用について, マルチメディア, 分散, 協調とモバイル (DICOMO 2006) シンポジウム, pp. 777–780 (2006).

[4] 柴田賢介, 荒金陽助, 塩野入理, 金井敦: 電子メールの解析によるフィッシングおよびファーミング対策方式の提案, マルチメディア, 分散, 協調とモバイル (DICOMO 2006) シンポジウム, pp. 477–480 (2006).