RL-002

サイトの移動判定によるフィッシングサイト検知手法の検討

An Examination of a Detection Method of Phishing Site by Decision of Site Movements

五島 優美† 岡崎 直宣‡
Yumi Goto Naonobu Okazaki

1. はじめに

近年インターネットの普及により，利便性が高まる一方，個人情報の搾取を目的としたフィッシングという犯罪が社会問題となっている．

フィッシングとは，オンライン上で不正に個人の機密情報が搾取されることである．一般的にフィッシングと言われるものは，その中の詐欺フィッシングである[1]．この詐欺フィッシングとは，フィッシング詐欺師（フィッシャー）が金融機関などの正規のメールを装い，個人情報の入力を促す電子メール（フィッシングメール）を不特定多数に送りつけ，それを受け取ったユーザにそのフィッシングメール内に記載されているリンク先（フィッシングサイト）で個人情報を入力，送信させることによって，個人情報を搾取するものである．

フィッシングは米国で大規模な被害を出している．日本国内においてもオンラインサービスの定着と共にフィッシングによる被害が出てきており，被害拡大の前に早急な対策が求められる．

フィッシングサイト検知手法の既存の対策手法としてホワイトリスト方式，ブラックリスト方式がある．ホワイトリスト方式とは，正規サイトを登録したホワイトリストと判定する対象のサイトを比較し，そのリストに載っていないサイトを疑わしいサイトとしてはじく方式である[2]．ブラックリスト方式とは，過去に発見されたフィッシングサイトを登録したブラックリストと判定する対象のサイトを比較し，リストに載っているサイトをフィッシングサイトとしてはじく方式である[3]．しかし，このホワイトリスト方式とブラックリスト方式を用いても判断することのできないグレーゾーンのサイトが存在してしまう．

本論文では，フィッシングサイトにおける個人情報の搾取を阻止するとともに，フィッシングサイトを検知することを目的とし，個人情報搾取後に正規サイトに移動するという特徴をもつフィッシングサイトに対して，擬似情報送信後の対象サイトの動向によって検知する手法を提案する．

2. フィッシング

2.1 日本のフィッシング

フィッシング対策協議会[4]によると，国内での最近のフィッシング事例としては，2008 年 3 月にゆうちょ銀行をかたったフィッシングサイトがある．これは，重要なお知らせがあるとして，偽サイトに誘導しようとしたものである．また，2008 年 2 月には mixi.jp の体裁を真似た「mixii（ミクスィ）」というサイトが確認された．これは，mixi の新規登録画面に似せて作られており、メールアドレスや年齢などを入力するよう促したものである．

†宮崎大学大学院工学研究科情報システム工学専攻
‡宮崎大学工学部

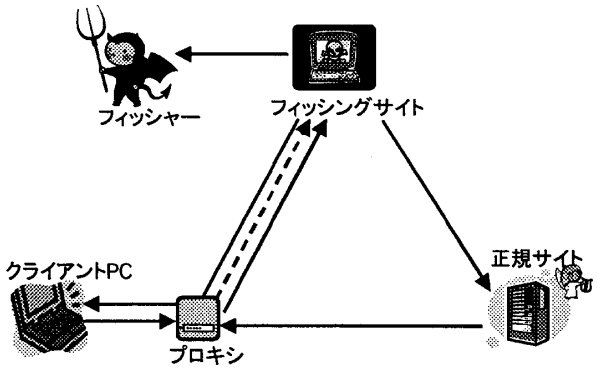

図 1 ネットワーク構成

2.2 フィッシングサイトの特徴

フィッシングサイトには，平均残存期間が非常に短いことなど独自の特徴がある．

その他，フィッシングサイトは自らの存在を隠し，ユーザに自らを正規サイトであると思わせるという性質がある．このため，フィッシングサイトは正規サイトが視覚的に類似していることや，ログインページなど個人情報を搾取するためのページ以外は正規サイトに移動するなどの特徴もある．

3. 提案手法

フィッシングサイトはユーザに自らの存在を隠そうとするため，個人情報搾取ページ以外は正規サイトへ移動するという特徴をもつものがある．そのため，ここではログインページ以外は正規サイトへ移動するフィッシングサイトを対象とする．

フィッシングサイトはユーザに自らを正規のサイトであると信じ込ませるために正規のサイトであっても起こりうるような動作をする．そのため，口座情報等の搾取を目的としている場合でも，多くの場合，ログインページからログインをさせ，認証したふりをして口座情報入力ページへ移動させると考えられる．そのため，ここでは搾取される個人情報は，ログイン情報である ID とパスワードとする．

本手法の前提条件としては，判定対象サイトはホワイトリスト方式やブラックリスト方式を用いて判断することができなかったグレーゾーンのサイトであるとする．また，ログイン情報は英数字のみとし，SSL(Secure Socket Layer)暗号化のサーバ証明書は偽造されていないものとする．

本手法は，ユーザの個人情報の送信を契機として動作し，ユーザが入力した個人情報を元に作成した擬似情報を送信する．そして，対象サイトから他のサイトへ移動するかどうかを観察する．また，ユーザが個人情報を入

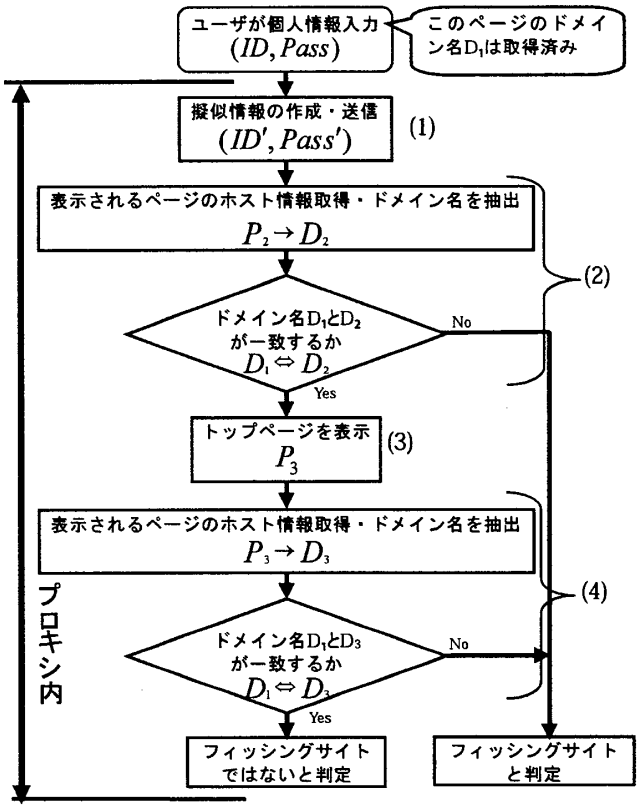

図 2　処理手順

力し，送信しようとした後の本手法の動作はすべてプロキシ内で行われるものとし，ユーザからのアクションはないものとする．このプロキシはユーザが閲覧している Web ページのパケットを常に監視しているものとする．図 1 は，本手法におけるネットワーク構成の図である．

3.1　処理手順

クライアント PC と Web サーバの間に設置されたプロキシは，常に表示される Web ページのホスト情報を取得し，ドメイン名を一時的に保持しているものとする．従って，ユーザが個人情報を入力した時点で既にその個人情報入力ページのドメイン名は取得されているものとする．

次に，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（図 2）．

(1) ユーザが入力した個人情報(ID,Pass)を元に擬似情報(ID',Pass')を作成し，ユーザの入力した個人情報(ID,Pass)の代わりに作成した擬似情報(ID',Pass')を送信する．

(2) 擬似情報(ID',Pass')を送信した後に表示されたページ P_2 のホスト情報を取得し，ドメイン名 D_2 を抽出する．このドメイン名 D_2 と最初に表示された個人情報入力ページ P_1 のドメイン名 D_1 が一致するかどうか比較し，サイトの移動判定を行う．
- ドメイン名 D_1 と D_2 が一致しなかった場合，サイトが移動したと判定し，このサイトをフィッシングサイトと判定する．
- ドメイン名 D_1 と D_2 が一致した場合，同一のサイト内のページであると判定し，次のステップへ進む．

(3) サイトの移動が確認されなかった場合，個人情報入力ページ以外のページ（トップページ等）P_3 へ移動する．

(4) 表示した個人情報入力ページ以外のページ P_3 のホスト情報を取得し，ドメイン名 D_3 を抽出する．このドメイン名 D_3 と最初に表示された個人情報入力ページ P_1 のドメイン名 D_1 が一致するかどうか比較し，サイトの移動判定を行う．
- ドメイン名 D_1 と D_3 が一致しなかった場合，サイトが移動したと判定し，フィッシングサイトと判定する．
- ドメイン名 D_1 と D_3 が一致した場合，同一サイト内のページであると判定し，フィッシングサイトではないと判定する．

3.2　擬似情報作成

3.2.1　個人情報の特徴

擬似情報作成のため，大手ポータルサイトやネットバンクなどの ID とパスワードについて調査した．

- ID
  - 小文字の半角英数
  - 桁数はサイトによって様々
  - 最初はアルファベットなどの制限がある場合がある
  - メールアドレスの場合がある
  - 金融機関は口座番号の場合がある
- パスワード
  - 半角英数
  - 大文字と小文字の区別がある
  - 記号が使われる（サイトによって種類は違う）
  - スペースも使われる場合がある

3.2.2　擬似情報作成

本手法では，フィッシングサイトに擬似情報を正しい個人情報と信じ込ませるため，正規サイトでの在り得るような擬似情報を作成する．

ID とパスワードについての調査結果から，擬似情報作成には次のようなポイント挙げられる．

- 桁数
  - サイトによって有効な桁数が異なるため，ユーザの入力した個人情報から桁数を取得する必要がある．
- 文字の種類
  - 半角英数でも大文字と小文字の区別や，最初の文字はアルファベットなどの制限があるため，文字の種類とその位置を判別しなければならない．
- 記号
  - 有効な記号はサイトによって様々であり，擬似情報作成においてもそのサイトにおいて有効な記号を用いる必要がある．
- メールアドレス
  - メールアドレスを ID として使用するサイトも存在し，サイトでメールアドレスのドメインが指定されている場合もある．そのため，ID がメールアドレスであるか判断する必要があり，さらにメールアドレスのドメイン部は変更せずに擬似情報を作成しなければならない．

このような擬似情報作成においてのポイントを考慮し，擬似情報を作成する．

以下に擬似情報の作成手順を示す．ここでは ID の擬似情報作成を例に説明をする．なお，パスワードの場合も同様の手順で行う．

(1) ユーザの入力した個人情報 ID の桁数 n をスペースも含めてカウントする．
(2) 一文字ずつ文字の種類（小文字英字，大文字英字，数字，記号）を判断する．
(3) 判定した文字の種類が英数字の場合，乱数を用いて同じ種類の違う文字へ変換する．ただし，記号及び@以降の文字列は変換を行わない．
(4) (2), (3)を個人情報の桁数 n の回数分繰り返す．
(5) 擬似情報の文字列 ID'とユーザの入力した正規情報の文字列 ID を比較し，違うものになっていることを確認する．擬似情報 ID'と正規情報 ID が違うもの（ID ≠ID'）になっていた場合，擬似情報 ID'の完成である．同様の文字列が擬似情報として作成されていた（ID= ID'）場合には，もう一度，新たな擬似情報の作成を行う．

3.3 サイトの移動判定

本手法のサイト移動の判定にはドメイン名を利用する．個人情報を送信しようとしたページのドメイン名とその後表示されるページのドメイン名を比較し，ドメイン名が一致すれば，同一サイト内のページであり，サイトの移動はないと判定する．ドメイン名が一致しなかった場合には，サイトの移動が行われたと判定する．

3.3.1 ホスト情報の取得

本手法では，表示ページのホスト情報を取得する方法は，A)HTTP による通信と，B)SSL による通信のどちらで通信が行われているかで２通りに分ける．

Web ページが変わるときはユーザによるクリック等の何らかのアクションがあるものとして，アクション後の説明をする．また，WWW は TCP を利用しているため，TCP による通信のみに着目する．

A) アクション後の最初に HTTP による通信が行われていた場合
  (1) HTTP リクエストを順序通り保持しておく．
  (2) それぞれのリクエストのレスポンスの content type をチェックし，それが text/html 以外のものは棄却する．（図 3-(a)）
  (3) content type が text/html である HTTP リクエストの中の一番初めの HTTP リクエストとレスポンス内にメッセージヘッダの location があるかどうかチェックする．（図 3-(b)）
  (4) location があった場合，移動先への HTTP リクエストとレスポンスに着目し，最終目的の HTTP リクエストを選択する．
  (5) location がなければ，その一番初めの HTTP リクエストを選択する．
  (6) 選択した HTTP リクエストからホスト情報を取得する．（図 4-(c)）
B) アクション後，HTTP リクエストより先に SSL による通信が行われていた場合
  (1) HTTP リクエストより先に HTTPS で通信しようとした場合，接続を確立した相手先の IP アドレスを取得する．また，ログインページで擬似情報送信後に再ログインページが表示された場合など既に HTTPS の接続が確立されていた場合には，ACK を出した相手先の IP アドレスを取得する．
  (2) 取得した IP アドレスから DNS で逆引きをする．
  (3) 逆引きできた場合，その逆引き結果をホスト情報として取得する．
  (4) 逆引きできなかった場合，SSL のセクションを確立したときに得られる SSL サーバ証明書から，その証明書の発行先をホスト情報として取得する．（図 5-(d1)，図 5-(d2)）

```
HTTP/1.1 302 Found
Date: Web, 01 Jan 2008 00:00:00 GMT
P3P: CP="NON CUR OTPi OUR NOR UNI"
Cache-Control: no-cache
Location: http://xxx.ab.co.jp -(b)
Connection: close
Content-Type: text/html -(a)
```

図３ HTTP レスポンスの例

```
GET /r/l1 HTTP/1.1
Accept: image/gif, image/jpeg, ・・・
Referer: http://xxx.xxx.co.jp
Accept-Language: ja
UA-CPU: x86
Accept-Encoding: gzip, deflate
User-Agent: Mozilla/4.0
Host: xxx.xxx.co.jp  -(c)
Connection: Keep-Alive
Cookie: A=aaaaaallaaaa; ・・・
```

図４ HTTP リクエストの例

```
TLSv1 Record Layer: Handshake Protocol:
Certificate
//省略//
 subject:rdnSequence
  rdnSequence:6items
   Item:litem(id-at-countryName=JP)
   Item:litem
 (id-at-stateOrProvinceName=Miyazaki)
   Item: 1 item (id-at-
localityName=Miyazaki-shi)
   Item:litem(id-at-organizationName=XXX
Corporation)
   Item:litem
(id-at-organizationalunitName=Loginxxx)
  Item:litem(id-at-commonNme=login.xxx.co.jp)
  Item (id-at-commonNme=login.xxx.co.jp)-(d1)
    DirectoryString: printablestring
      printablestring: login.xxx.co.jp -(d2)
//省略//
```

図5 サーバ証明書の例

3.3.2 ドメイン名の抽出

本手法のドメイン名抽出方法は，gTLD か ccTLD で取得するドメイン名のレベルが変わる．また日本の ccTLD の JP ドメイン名の場合は，SLD で JP ドメイン名の種類を判断し，さらにその種類によって変化する．以下で，ドメイン名を抽出する手順を，JP ドメイン名を例に説明する．

(1) TLD から，gTLD か ccTLD かを判断する．
(2) gTLD の場合，「SLD + gTLD」を取得する．ただし，新 gTLD の name，pro は 3rd level domain まで取得する．
(3) ccTLD の場合，SLD から JP ドメインの種類を判断する．
  - 汎用 JP ドメイン名の場合「SLD + ccTLD」を取得する．
  - 属性型 JP ドメイン名の場合「3rd level domain + SLD + ccTLD」を取得する．

- 地域型 JP ドメイン名の場合，さらに 3rd level domain から，一般地域型ドメイン名か地方公共団体ドメイン名か判断する．
  - ➢ 一般地域型ドメイン名の場合，「4th level domain + 3rd level domain + SLD + ccTLD」を取得する．
  - ➢ 地方公共団体ドメイン名の場合，「3rd level domain + SLD + ccTLD」を取得する．

3.4 ページの移動

本手法では擬似情報を送信後サイトの移動が確認できなかった場合，個人情報入力ページ以外のページ（トップページ等）を表示し，そのページが個人情報入力ページと同じサイト内にあるかどうかで，そのサイトがフィッシングサイトかどうかを判定する．

多くのサイトでは，どのページからもそのサイトのトップページへのリンクが張られているため，HTML 解析を行い「トップ」，「ホーム」等の単語でリンクを張られている先を表示することでトップページが表示できると考えられる．

また，「トップ」，「ホーム」等の単語がない場合，ログイン画面にほとんどあると考えられる「初回登録」，「はじめてのログイン」，「パスワードを忘れた方」，「ログインできない方」等の単語で張られているリンク先を表示する．

さらに，これらのような単語でリンクが張られていなかった場合には，そのページに張られているリンクの中でページの一番上に張られているリンクのリンク先を表示する．

4.評価

ネットワーク・アナライザ・ソフトである WireShark[5]を用いて，パケットを収集し，ホスト情報を取得した．そのホスト情報からドメイン名を抽出し，ドメイン名が一致するかどうかによってサイトの移動判定を行い，本手法のフィッシングサイトに対する検知能力の評価と，正規サイトに対する誤検知の評価を行った．

4.1 フィッシングサイトに対する評価

本手法のフィッシングサイト検知能力の評価のため，PhishTank[6]に登録されているフィッシングサイト 20 サイトについて評価を行った．

20 サイトのうち，本手法が対象としているような個人情報搾取後に正規サイトへ移動するフィッシングサイトやトップページ等の個人情報搾取ページ以外は正規サイトへ移動するようなフィッシングサイトは 13 サイトであった．この 13 サイトにおいて，本手法のフィッシングサイト検知能力を評価した．

13 サイト中フィッシングサイトであると正しく判定されたサイトが 12 サイトであり，1 サイトは検知することができなかった．

この 1 サイトは，IP アドレスのままのサイトであり，ドメイン名を取得することができなかったため，サイトの移動判定ができず，検知することができなかった．

対策として，HTTP リクエスト内の Host 情報の部分が IP アドレスだった場合には，IP アドレスで比較することなどが考えられる．

また，本手法では対象としないフィッシングサイトと判断した 7 サイトは，個人情報搾取ページ以外のページへのリンクがないサイトや，リンク先へ移動できないサイト，サイト内の全てのページをフィッシングサイト内においているサイトであった．

4.2 正規サイトに対する評価

正規サイトを誤検知してしまう可能性についての評価を行うため，国内銀行や大手ポータルサイトなど正規サイト 30 サイトについて，本手法で表示されると考えられるログインページや Top ページのドメイン名を取得し，同一サイト内でドメイン名が一致するかどうか検証した．

同一サイト内でドメイン名がすべて一致して正しく判定されたサイトは 30 サイト中 27 サイトとなった．

ドメイン名が一致しないページが存在したサイトは，ネットワークバンキング共同センササービス等の外部サービスを導入している場合の他に，複数のドメイン名がサイト内で使い分けられている可能性がある．

複数のドメイン名がサイト内で使い分けられている場合の対策として，WHOIS[7]などのドメイン名登録情報検索サービスを使用し，ドメイン名の所有者を取得し，所有者が同一のドメイン名であればサイトが移動していないと判定するなどが考えられる．これによって 1 つの企業が複数のドメインを取得し，同一サイト内で使用していた場合であっても，異なるサイトと判定せずに同一サイトであると判定することが可能となる．

5. まとめ

本論文では，ホワイトリスト方式やブラックリスト方式で検知することのできないグレーゾーンに分類されるサイトにおいて，擬似情報を送信し，その後の判定対象サイトの動作によって，個人情報を守りつつ，フィッシングサイトを検知する手法を提案した．

提案手法は，対象とするようなサイトの移動を伴うフィッシングサイトには有効であると考えられる．

今後は，提案手法の実装を行い，より多くののフィッシングサイトに対する評価を行い，実用性を検証したい．

今後の課題として，ホスト情報を取得することのできない IP アドレスのままのフィッシングサイトへの対応や，ID やパスワード以外の個人情報搾取ページへの対応などが挙げられる．

参考文献

[1]Aaron Emigh, Radix Labs, “Online Identify Theft: Phishing Technology, Chokepoints and Countermeasures”, http://www.antiphishing.org/Phishig-dhs\report.pdf

[2]柴田賢介, 荒金陽助, 塩野入理, 金井敦, “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”, 情報処理学会 DPS Vol.2006 No.96 pp.17-22

[3]RBL.JP, http://www.rbl.jp

[4]フィッシング対策協議会, http://www.antiphishing.jp/

[5]WireShark,http://www.wireshark.org/

[6]PhishTank, http://www.phishtank.com/

[7]JP WHOIS /JPRS, http://whois.jprs.jp/