6V-08

コンテンツベースフィッシング検知方式の誤検知率改善の試み

浜地　達也†　　岡本　剛†
神奈川工科大学†

1. はじめに

フィッシング詐欺とは，正規のサイトを模倣した，フィッシングサイトで情報窃取をする行為である．主な手口としては，攻撃者がメールで金融機関などの組織を名乗り，フィッシングサイトに誘導する．そして，誘導先で個人情報を入力させる．フィッシング詐欺を減らすためには，詐欺で用いられるフィッシングサイトの検知率の向上が必要である．そこで本研究では，フィッシング検知技術の一つである，コンテンツベース方式の誤検知率改善を試みる．

2. コンテンツベース検知方式

コンテンツベースのフィッシングサイト検知方式は，フィッシングサイトが正規サイトと類似したコンテンツを持ち，そのコンテンツに含まれる単語を Web 検索したとき，検索結果の上位には，正規サイトがヒットすることに基づく．Web 検索結果の上位サイトが正規サイトと判断する根拠は，フィッシングサイトの存在期間が平均 3.1 日と短く，検索エンジンからの評価が低いという事実に基づく．コンテンツベースのフィッシングサイト検知方式の処理手順は図 1 の通りである．

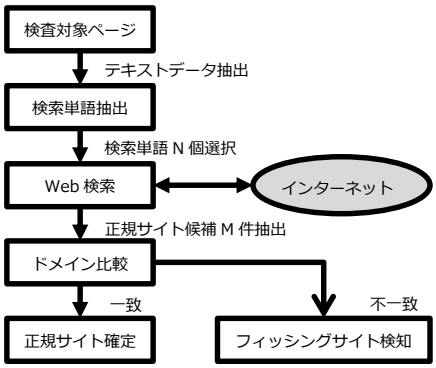

図 1　コンテンツベース方式の処理の流れ

検査対象のページからテキストデータを抽出したら，特徴的な単語を抽出するため，TF-IDF により単語の重み付けを行う．TF-IDF の重み付けは以下の式から計算する．

$$\mathrm{tfidf} = \mathrm{tf} \times \mathrm{idf}$$

$$\mathrm{tf(t,d)} = \frac{n_{t,d}}{\sum_{s \in d} n_{s,d}}$$

$$\mathrm{idf(t)} = \log \frac{N}{df(t)} + 1$$

tfは単語の出現頻度であり，ページ内での特徴度を表す．idfは逆文書出現頻度であり，世界中の Web 全体での特徴度を表す．$n_{t,d}$はページ内での単語 t の出現回数を表す．$\sum_{s \in d} n_{s,d}$はページ内での単語総数を表す．Nは全世界の Web サイト総数である．本研究では 2013 年の Google のサイトを参考に 60 兆とした．df(t)は単語 t の Google 検索ヒット数を表す．これらの式により，他のサイトに含まれないページ特有の単語の重み付けを行う．TF-IDF 値の大きい単語の上位 N 個を検索単語とする．検索結果の上位 M 件のサイトのドメインと調査対象ページのドメインを比較し，一致したら正規サイトとし，不一致ならフィッシングサイトと判定する．

文献 [1]では，英字を利用する言語を対象としていたので，本研究では，次に示す方法により日本語に対応させる．抽出したテキストデータを形態素解析ソフト MeCab により単語に分割する．ここで抽出する単語は，品詞が「名詞」であり，品詞細分類が「一般」または「固有名詞」であるものとする．そして，各単語の出現回数をカウントする．これを TF-IDF により単語の重み付けを行う．

2.1. コンテンツベース方式の問題点

コンテンツベース方式は，先行研究 [2]により次の問題点が指摘されている．

1) 認証ページ特有の「パスワード」などの単語が多数抽出され，TF-IDF が有効に機能しない．
2) 画像が多く，テキストが少ないサイトでは，特徴的な検索単語を抽出できない．
3) 同じ組織のサイトでも，ページによって複数のドメインで運用されているため，ドメインが一致しないことがある．

3. 誤検知率改善の提案手法

コンテンツベースのフィッシングサイト検知方式の問題点を改善するために次の手法を提案する．

Approaches for Reducing False Positives on Content-Based Phishing Detection
†Tatsuya Hamachi, Takeshi Okamoto, Kanagawa Institute of Technology

1) 認証ページで共通に出現する単語を検索単語から除外する.
2) 検索単語を増やすために，画像データからテキストを抽出する.
3) Whois 情報の登録事業者を参照する.

本研究ではテキスト抽出に Google Chrome を使用する．Google Chrome の拡張機能である AdBlock を利用して，サイトの広告を表示させないようにした.

3.1. 共通単語除外の方法

TF-IDF による特徴的な単語抽出を行っても，「パスワード」などのようなサイトの特徴を示さない単語が選ばれることがある．そこで，Alexa の日本のアクセスランキングから，日本語の認証ページのあるサイト 100 件から，各サイトに共通する単語を抽出した．抽出方法は，1 サイトに出現した単語を，サイト内の出現回数にかかわらず 1 カウントとして，カウントの多い単語から降順に上位から 30 単語を共通単語とした．30 単語にした根拠は，これ以上単語を増やした場合に，「google」などのサイト名が選ばれるからである．30 個の共通単語を TF-IDF の名詞抽出後に除外した.

3.2. 画像テキスト化の方法

評価対象ページの画像から手動でテキストを抽出し，検索単語の候補を増やす．この手法を行う理由は，OCR を利用することにより，誤検知率が改善されるかを調査である．そのため，OCR による読み取りが難しいとされる，以下の条件を除いたテキストデータを抽出した.

1) ロゴマークなどの装飾が付いたフォント
2) テキストの背景に画像があるもの

3.3. Whois 情報参照の方法

図 1 のドメイン比較において，不一致であった場合に Whois 情報参照の処理を追加する．Whois 情報を参照し，Organization（jp）や Registrant Organization（com，org）などの登録事業者を比較する．例えば Microsoft 関連サイト(microsoft.com，xbox.com など)と，その認証ページ(live.com)はドメインが異なる．このとき，認証ページ以外が有名なため，検索結果上位に認証ページ以外のドメインが出現する．そこで，サイトを運用している事業者を比較することにより，誤検知を減らせると考えた.

4. 提案手法の有効性評価

提案手法の有効を確認するため，Alexa の日本のアクセスランキングから，日本語の認証ページがあるサイト 100 件を有効性評価の対象とした．認証ページで評価を行う理由は，フィッシング詐欺により，認証ページで個人情報を窃取されるためである．本研究では，正規サイトの誤検知率とフィッシングサイトの検知率を評価した．コンテンツベースのフィッシングサイト検知方式のパラメータである検索単語数 M は，3，5，及び 10 個とし，検索結果件数 N を 10 とした．改善前のコンテンツベース方式を含む 4 種類の誤検知率を図 2 に示す.

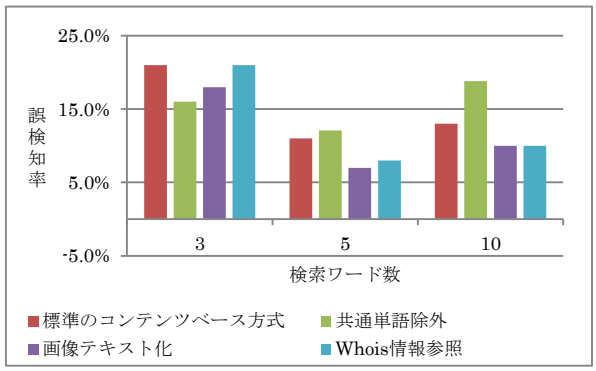

図 2 各手法の誤検知率

図 2 の結果から，画像テキスト化と Whois 情報参照は，改善前のコンテンツベース方式と比較して，全体的に誤検知率が減少した．一方，共通単語除外は検索単語 5 個で 1.1 ポイント，10 単語で 5.8 ポイント誤検知率が増加した．しかし，検索単語 3 個の状態では他の手法と比べて，最も誤検知率が減少していた．提案手法を組み合わせた場合でも，共通単語除外を含む組み合わせは，改善前のコンテンツベース方式と比べて，検索単語が少なければ誤検知率が減少し，多ければ増加した．画像テキスト化と Whois 情報参照の組み合わせでは各手法単体よりも誤検知率が減少した．以上から，画像テキスト化と Whois 情報参照は誤検知率の減少に有効である．また，共通単語除外は少ない検索単語で誤検知率を減少する際に有効である.

5. おわりに

コンテンツベースフィッシング検知方式は，正規サイトをフィッシングサイトであると誤検知する割合が高いため，まだ実用的ではない．本研究の提案手法により，正規サイトをフィッシングサイトであると誤検知する割合を最大で 61.9%改善した.

6. 参考文献

1) Yue Zhang，et al.: CANTINA: A Content-Based Approach to Detecting Phishing Web Sites，2007.
2) 松岡孝幸：日本語 Web サイトに対するコンテンツベースフィッシング検知方式の有効性評価，神奈川工科大学　卒業論文，2014.