6ZD-06

オリジナルサイトとフィッシングサイトの定量的類似度評価

角田 晴輝　寺田 真敏
東京電機大学

1. はじめに

フィッシングサイトによる被害は増加傾向にあり，社会的影響も大きい．本研究は，フィッシングサイトがオリジナルサイトと見た目でほとんど差異がないよう巧妙に構築されていることに着目し，その類似性を複数の視点から定量的に評価することを通して，対策のための知見蓄積を目的としている．

本稿では，オリジナルサイトとフィッシングサイトにアクセスした時に取得できるデータから見た目はどの程度似ているのか，どのように差異がないよう作成しているのかなど数値化した調査結果を報告する．

2. 関連研究

フィッシングに関する論文 46 件を 4 つの研究区分に分類した結果を表 1 に示す．文献[1][2]など，フィッシングサイトの分析に関する研究は数多く取り組まれているが，フィッシングの被害の原因として挙げられるオリジナルサイトと差異なく巧妙に作成されていることに着目し，定量的評価などを通してどれだけ似ているかの研究は実施されていないことを確認した．

表 1：フィッシングに関する研究区分

区分	概要	件数
状況把握	状況把握から対策の提案や考察	2 件
分析	フィッシングサイトの攻撃実態	3 件
	対策・検知システム	4 件
検知	ページ遷移	2 件
	機械学習	2 件
	URL	6 件
	コンテンツベース	3 件
	HTTP リクエスト	2 件
	画像	4 件
	その他	4 件
対策	ユーザ行動	7 件
	セキュリティ教育	3 件
	その他	4 件

3. 類似度調査

3.1 調査目的

本調査では，次に示す 3 つの観点から定量的類似度評価による調査をし，フィッシングサイトへの対策の知見蓄積を目的とする．

- 見た目の類似度
- サイト構成上の類似度
- ファイル構造上の類似度

3.2 調査対象とデータ収集

調査対象は，2022 年 3～11 月に電子メールで受信した 36 件のフィッシングサイトを対象とした(表 2)．調査に使用するデータは，オリジナルサイトとフィッシングサイトにアクセスし，表 3 に示すデータ形式で収集した．

表 2：調査対象サイト一覧

オリジナルサイトの分類	オリジナルサイト件数	フィッシングサイト件数
銀行系	3 件	7 件
カード系	8 件	21 件
その他	5 件	8 件

表 3：収集するデータ一覧

区分	概要
ウェブページ全体	html ファイルと関連ファイルを含む完全のデータ． ウェブブラウザの画面右上の「…」アイコン→「名前を付けてページを保存」で「ウェブページ，完全」で保存．
HTML ソースコード	html ファイルのみのデータ(view-source ファイル)． ページ上で右クリック→「ページのソースを表示」で「名前を付けた保存」を「ウェブページ，HTML」のみで保存．
スクリーンショット	jpg 形式 横 1440×縦 900 のサイズで保存．
HTTP 通信データ	har 形式 ページ上で右クリック→「F12」キー押下．表示されたパネルから「Network」タブ選択→「Disable cache」項目にチェックを入れて保存．

3.3 調査内容

調査項目の詳細を表 4 に示す．

表 4：調査項目一覧

見た目の類似度	
サイトのスクリーンショット	Average Hash による画像の類似度を算出． 両サイトのスクリーンショットの Average Hash の値の差を算出．差の値が 0 の場合を類似度 100，20 の場合を類似度 0 として類似度を算出． 計算方法： $\text{類似度} = 5 \times (20 - \text{両サイトのスクリーンショットの}\mathit{Averagehash}\text{の差})$
画像ファイル	両サイトで使用する同名の画像ファイルを対象に，データが完全一致する画像ファイル数から類似度を算出． 計算方法： $\text{類似度} = \frac{\text{ファイルの中身が同じファイル総数}}{\text{両サイトに存在するファイル総数}} \times 100$
サイト構成上の類似度	
css・js ファイル，画像ファイルの名称，ディレクトリパス	両サイトの view-source ファイルから，ファイル名・ディレクトリパスを抽出して類似度を算出．ディレクトリパスは部分一致で調査し，階層ごとの類似度を算出． 計算方法： (上段：通常，下段：フィッシングサイトのみ存在するファイルがある場合の方法) $\text{類似度} = \frac{\text{一致したファイル総数}}{\text{オリジナルサイトのファイル総数}} \times 100$ $\text{類似度} = \frac{(\text{一致したファイル総数}) - (\text{フィッシングサイトのみ存在するファイル総数})}{\text{オリジナルサイトのファイル総数}} \times 100$
オリジナルサイトの参照度	両サイトの view-source ファイルから，オリジナルサイトから直接参照するファイルの参照度を算出． 計算方法： $\text{類似度} = \frac{\text{参照したファイル総数}}{\text{オリジナルサイトのファイル総数}} \times 100$
ファイル構造上の類似度	
css・js ファイルの構造	両サイトで使用する同名の css・js ファイルから，行単位で構造の類似度を算出． 計算方法： $\text{類似度} = \frac{\text{ファイル名が一致した全}\mathit{css・js}\text{ファイルの一致した行の総数}}{\text{ファイル名が一致した全}\mathit{css・js}\text{ファイルの総行数}} \times 100$

Similarity evaluation between Original and Phishing Site
Haruki Kakuta and Masato Terada, Tokyo Denki University

4. 調査結果

4.1 見た目の類似度

一部サイトを除いて同じ画像を使用するサイトが多く(図 1 の下段), スクリーンショットでも類似度が高いサイトが多いことから見た目が酷似しているサイトが多いことが分かる (図 1 の上段).

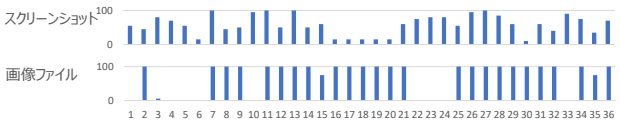

図 1：見た目の類似度

4.2 サイト構成上の類似度

(1) css・js ファイルの名称，ディレクトリパスの類似度

名称は類似度の高いサイトが多いことから，オリジナルからファイルをダウンロードしていることが分かる(図 2 の 1, 2 段目). ディレクトリパスはファイル構成を部分的に模倣しているか判断するために階層毎に調査し，その結果一部を除いて類似度が 0 に近いサイトが多いことから，ファイル構成までは模倣していないことが分かる(図 2 の 3～8 段目).

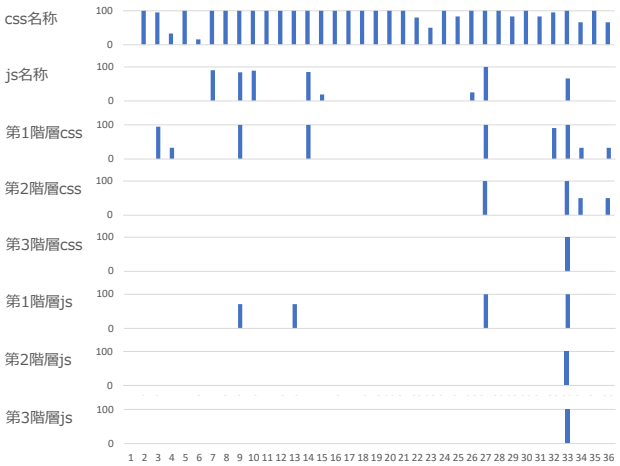

図 2：css・js ファイルの名称，ディレクトリパスの類似度

(2) 画像ファイルの名称，ディレクトリパスの類似度

css・js と同様に名称はオリジナルサイトからファイルをダウンロードしていることが分かる(図 3 の 1 段目). ディレクトリパスは一部のサイトを除いて類似度が 0 に近いサイトが多いことから，画像に関してもファイル構成は模倣していないことが分かる (図 3 の 2～4 段目).

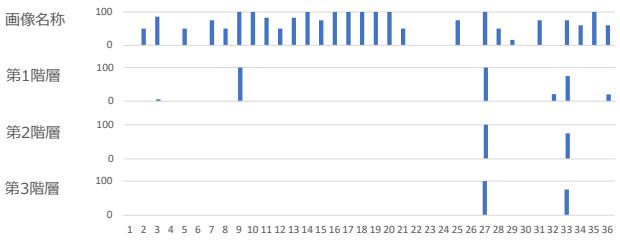

図 3：画像ファイルの名称，ディレクトリパスの類似度

(3) オリジナルサイトの参照度

css・js，画像のいずれも，フィッシングサイト側で該当ファイルを保持していることが分かる(図 4).

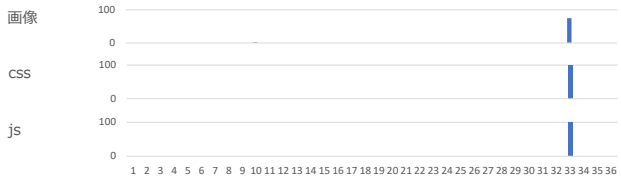

図 4：オリジナルサイトの参照度

4.3 ファイル構造上の類似度

css は類似度が高いサイトが多く(図 5 の上段)，js は類似度が低いサイトが多い(図 5 の下段).

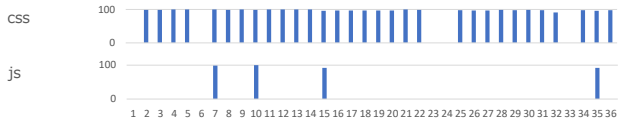

図 5：ファイル構造の類似度

4.4 考察

(1) 見た目の類似度

スクリーンショットを業種別に見ると(図 6)，銀行系，カード系の類似度が低く，その他が高い．銀行系やカード系のサイトでは ID パスワード以外の認証方法(パズル認証や CAPTCHA 機能)，フィッシング等の注意喚起などを掲載しており，フィッシングサイトでは削除されていることに関係している．

(緑：銀行系，黄：カード系，赤：その他)

図 6：業種別スクリーンショットの類似度

(2) サイト構成上の類似度

見た目でユーザを騙すため，css や画像は同名のファイルを利用し，見た目に必要ない js の多くが削除されていると考える．しかし，ディレクトリパスでは完全と部分一致でも類似度が高くないことから，オリジナルサイトを模倣せず作成していると考えられる．

(3) ファイル構造上の類似度

css の類似度は高く，js は低い傾向にある．また，ファイル構造上同一と言えないサイトが多いことから，オリジナルサイトからコピーするだけでなくコードを細かく書き換えて使用していると考えられる．

5. まとめ

本稿では，フィッシングサイトへの対策の知見蓄積を目的としてフィッシングサイトの定量的評価の調査をした．今後，本稿で示した収集データをデータセットとして作成し，フィッシングサイト対策の新たな知見蓄積に繋げたいと考える．

参考文献

[1] 小寺 博和，小出 駿，千葉 大紀，青木 一史，秋山 満昭．偽ショッピングサイトを起点とする攻撃の実態調査．IPSJ SIG Technical Report．2020．Vol.2020-DPS-182 No.41.

[2] 高橋 啓伸，小倉 加奈代，Bhed Bahadur Bista，高田 豊雄．画像局所特徴量を利用したフィッシングサイト検知手法の実装と評価．Computer Security Symposium．2016．1234－1239.