DAY2　B2-5
アカデミック領域における
フィッシング詐欺対策研究の最新動向と研究事例紹介

ホモグラフドメインの検知に関する研究

2022/11/8

長崎県立大学
地域創生研究科 情報工学専攻

小嶋彬彦

目次

1. 背景

- フィッシングメールによる被害が後を絶たない
  - フィッシング詐欺の報告件数は右肩上がりとなっており
    月に10万件を超えている
  - 2022年には国税庁や金融庁などの公的機関を騙る
    フィッシングメールが報告されている

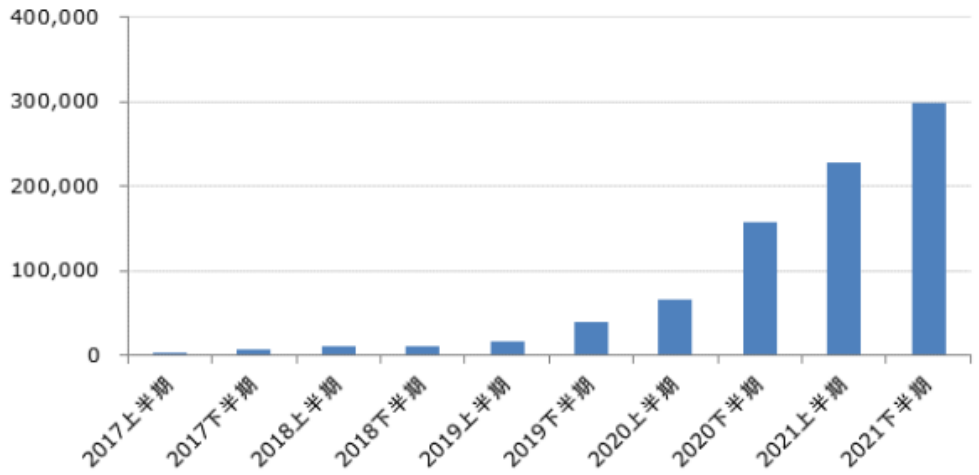


図 1-1　国内のフィッシング情報の届け出件数[2]

2. フィッシングメールの手法

- 攻撃者は以下のような手法で送信者のなりすましを行う
  - 送信者名・件名の偽装
    - メールソフトなどで表示される送信者名・件名を偽装する
  - 本文中に記述した内容によるなりすまし
    - メール本文になりすまし対象しか知りえない情報を記載する
  - メールアドレスの偽装
    - ユーザ名部分(@以前)をなりすまし対象に関連する文字列(会社名，名前，生年月日等)に設定
    - メールアドレスを**なりすまし対象のアドレスに酷似させる**

→ホモグラフドメイン

参照:https://www.soumu.go.jp/main_sosiki/joho_tsusin/security/business/staff/05.html
:https://www.ipa.go.jp/files/000009375.pdf

3. ホモグラフドメインとは

- ホモグラフドメインとは
  真正なドメインに酷似したドメイン名である
- 下のように人間の目での識別は困難である場合も多い

真正なドメイン	microsoft.com	google.com
ホモグラフドメイン	rnicrosoft.com 先頭文字がm（エム）ではなくrn（アールエヌ）	google.com 真正なドメインのeは文字コード0x65で、ホモグラフドメインのeは0xd0b5

3.1 ホモグラフドメインの現状

- 有名な企業や金融機関系のドメイン名では
  既に第三者が非常に似ているドメイン名を取得している
  ことが確認されている

- 2006年にはGoogle.comのホモグラフドメインは6個報告されていたが2017年には120個報告※されている(**20倍**)

※It’s Not What It Looks Like: Measuring Attacks and Defensive Registrations of Homograph Domains
Florian Quinkert, Tobias Lauinger, William Robertson, Engin Kirda, and Thorsten Holz
(https://ieeexplore.ieee.org/document/8802671)より

4. 先行研究(1)

- **OCR(光学的文字認識)** を用いたホモグラフドメインの検知手法
  澤部祐太、千葉大紀，秋山満昭，後藤滋樹，"OCRを利用したホモグラフIDNの検知法"，
  Computer Security Symposium 2018，2018

- **ドメインのwhois情報やSSIM値を用いた機械学習**による
  ホモグラフドメイン検知手法
  Hunting Brand Domain Forgery: A Scalable Classification for Homograph Attack
  ICT Systems Security and Privacy Protection 34th IFIP TC 11 International Conference
  Tran Phuong Thao, Yukiko Sawaya, Hoang-Quoc Nguyen-Son, Akira Yamada, Kazumasa Omote, and Ayumu Kubota 2019.

4. 先行研究(2)

- ホモグラフドメインをドメインの所有者が悪用防止で取得している場合もあるが，大多数は**取得されていない**

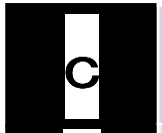
  Florian, Q., Tobias, L., William, R., Engin K. and Thorsten, H., "It’s Not What It Looks Like: Measuring Attacks and Defensive Registrations of Homograph Domains," 2019.

5. 提案手法

- 文字列の類似度を用いた検知手法を検討
- 本研究では文字列の類似度を計算する方法として
  ハッシュ関数の一種である**知覚ハッシュ(pHash)**を用いる
- ハッシュ値同士を比較し，**ハミング距離を類似度として扱う**

5.1 知覚ハッシュ

- マルチメディアデータ用の**ハッシュ関数の一種**
- **類似していると認識されたデータからは類似したハッシュ値が算出される**という性質を持つ

	文字(画像)	ハッシュ値(Hex)
	c	8fe0f00f0ff07887
	e	8fe0f00f0ff0f087
	A	aef0c30db49223bd

5.2 ハミング距離

等しい文字数を持つ2つの文字列の中で対応する位置にある
異なった文字の数

例:1234abと2235abのハミング距離は2

	文字(画像)	ハッシュ値(Hex)	cとのハミング距離
c	c	8fe0f00f0ff07887	0(同一)
e	e	8fe0f00f0ff0f087	2
A	A	aef0c30db49223bd	26

5.2 知覚ハッシュの種類

知覚ハッシュには複数の種類が存在する

- Average Hash(aHash)
  - 画像の平均輝度からの差分を用いる
- Perceptual Hash(pHash)
  - 画像を離散コサイン変換し，低周波数領域にaHashと同じ処理を行う
- Difference Hash(dHash)
  - 隣接領域との差分を用いる
- Wavelet Hash(wHash)
  - pHashの離散コサイン変換を離散ウェーブレット変換に変更したもの

5.3 各ハッシュ値の比較

以下の条件で各ハッシュ値を比較

- **大文字の「O」を基準**として文字画像から各文字の類似度を算出
- フォントはRoboto，比較用文字画像サイズは300px*300px
- 算出された類似度を主観と比較

	基準文字	最も類似度が高かった文字	ハミング距離(類似度)
aHash	O	U,W	1
pHash	O	0	2
dHash	O	0	3
wHash	O	R	1

5.4 提案手法と先行研究の比較

- OCRを用いる研究との比較
  - 検知に専用の環境が不要
  - 検知時に画像データを扱わないため処理量が少ない
- 機械学習を行う研究との比較
  - ラベル付けが不要で主観が影響しない
- ホモグラフドメインの定義の違い
  - 2文字対1文字(rn→mなど)で置換したもの含む

6. 評価結果

- 実際にフィッシングに使用された80のドメインリストを用いて評価
- 評価結果
  - 以下にあげたドメインの検知に成功
  - リスト内のホモグラフドメインを100%検知

真正なドメイン名	検知されたホモグラフドメイン
amazon.co.jp	anazom.co.jp
aeon.co.jp	acon.co.jp

7. 考察

- 提案手法により、ホモグラフドメイン以外のタイポスクワットや「.」部分を「-」に置き換えたドメインなどの中から**ホモグラフドメインを検知可能**
- 最も類似度が高い文字（列）を置換候補としたため**生成される候補数が少ない**
  - **どこまでなら類似していると見なすか要検討**
- **ASCII文字のみを対象にした**が，実際のホモグラフドメインは**非ASCII文字が用いられる事も多い**ため，非ASCII文字に対応する必要がある

8.まとめ

- フィッシングの被害が後を絶たない
- フィッシングメールに用いられるホモグラフドメインを知覚ハッシュを用いて検知する手法を提案
- 提案した手法で実際にホモグラフドメインを検知することが出来た
- 非ASCii文字への対応などで更なる検知制度の向上が見込まれる