ドメイン情報とHTTPレスポンスヘッダに基づくフィッシングサイトの識別と評価

桜井啓多†

明治大学総合数理学部 先端メディアサイエンス学科 菊池研究室†

1. はじめに

ユーザの情報を盗むフィッシングサイトの増加は大きな課題である．[1]によると，2017年下半期に届け出されたフィッシングサイトの件数は約5000件にのぼっている．

そこで，本研究では，本物と同一のコンテンツを持ち，人手では判別が困難なフィッシングサイトを検知する事を目的とする．ドメインとHTTPレスポンスヘッダの特徴に着目し，フィッシングサイトの情報を提供しているphishtank[2]内の100件のフィッシングサイトのドメイン情報とサイト側が設定しているHTTPレスポンスヘッダの情報の収集と分析をし，フィッシングサイトを自動検出する方法を提案する．

2 フィッシングサイトの特徴

2.1 収集した特徴の概要

収集したフィッシングサイトの主要な特徴量を表1に示す．表1の特徴量をフィッシングサイトと会社のホームページやWebサービス等の正規サイトを100件づつ求めた．

表1 ウェブサイトの特徴量

特徴	内容	例	重み
country	サイトを運用しているIPアドレスの国コード	US	0.4
domain interval	アクセス日時－ドメイン作成日時	98	0.45
domain lifetime	ドメイン期限日－ドメイン作成日時	6	0.45
x-xss-protection	HTTPレスポンスヘッダ．XSS攻撃を防止するための設定	True	0.33
x-frame-options	HTTPレスポンスヘッダ．クリックジャッキング攻撃を防止する設定	True	0.33
x-contetn-type-options	HTTPレスポンスヘッダ．コンテンツの内容を見ない様にする事でXSS攻撃のリスクを減らす設定	True	0.33
content-security-policy	HTTPレスポンスヘッダ．UAが読み込みを許可されたリソースを管理出来る様にするための設定	True	0.33

2,2 収集したドメインに関するデータ

サイトアクセス時にサイトのドメインが作成されてから，何日間経過しているかを表す"domain interval"の分布を図1に示す．その統計量を表2に示す．

ドメインが作成された日時からドメインの有効期限日までの期間を表す"domain lifetime"の分布を図2に示す．その統計量を表3に示す．

"domain interval"がドメインの新しさを示す指標であるのに対して，"domain lifetime"はドメインが運用される予定の期間を示している．

表2 domain interval

	正規サイト[日]	フィッシングサイト[日]
平均値	3089	135
中央値	3650	14
最頻値	3650	7

表3 domain lifetime

	正規サイト[年]	フィッシングサイト[年]
平均値	8.6	1.1
中央値	10	1
最頻値	10	1

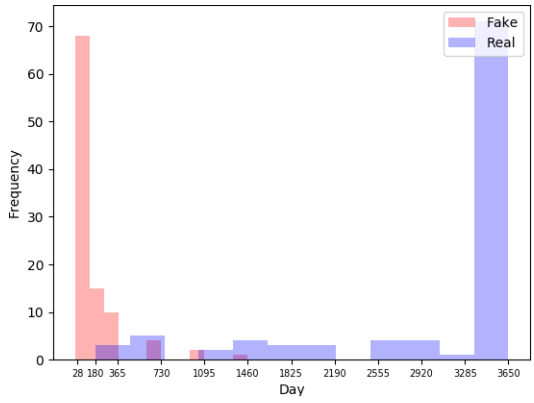

図1 domain intervalの分布

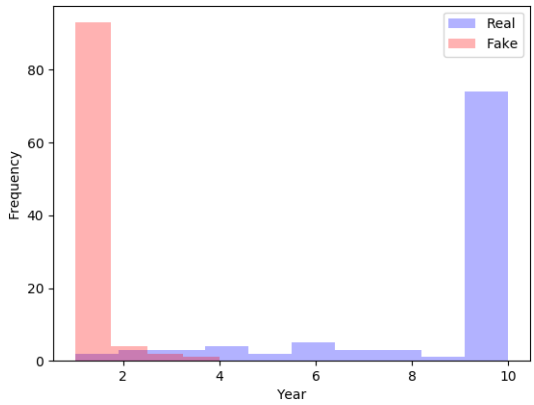

図2 domain lifetimeの分布

"domain lifetime"と"domain interval"は先行研究[3]を参考にした．いくつかの区間にカテゴライズされており10年以上の場合は10年として扱う．"domain interval"は正規サイトの最頻値が3650日であるのに対し，フィッシングサイトは7日と非常に期間が短い．"domain lifetime"も同様にフィッシングサイトの値は短い結果となった．

† Department of Frontier Media Science, School of Interdisciplinary Mathematical Science, Meiji University, Kikuchi Laboratory

2.3　HTTPレスポンスヘッダに関するデータ

HTTPレスポンスに関する項目は4つあり，正規サイトとフィッシングサイトで各々100件中設の件数を表4に示す．

表4 各HTTPレスポンスヘッダの件数

HTTPレスポンスヘッダ名	正規サイト	フィッシングサイト
x-frame-options	70	4
x-content-type-options	58	4
x-xss-protect	56	4
content-security-policy	25	2

2.4　考察

表4のHTTPレスポンスヘッダはユーザにセキュリティ情報を提供している．一方，フィッシングサイトはユーザの情報を盗むことが目的のサイトである点から，運用は短期間であり，セキュリティに関するヘッダーも設定されていないのではないかと考えた．表4は，明らかにこの仮説を裏付けている．

2.5　国に関するデータ

両サイトのIPアドレスの国分布を表5に示す．

表5 正規サイトの国分布

国コード	正規サイト	フィッシングサイト
US	56	40
JP	40	2
IE	2	0
TW	0	20
NL	0	6
UK	0	4
PA	0	4
RU	0	4
CA	0	3
KR	0	2
MA	0	2
FR	0	2
Other	2	9
Unknown	0	2

表5より，アメリカ(US)が最も多い事は共通だが，台湾(TW)，オランダ(NL)，ウクライナ(UK)，パナマ(PA)，ロシア(RU)などがフィッシングサイトのみで使用されている違いがある．

3　フィッシングサイト検出支援システム

2節で求めた特徴量を用いて，対象のサイトが安全であるかを判定するシステムをpythonで作成した．システム構成図を図3に示す．本システムはプロキシとして動作し，フィッシングサイトを検出した時にアクセスをブロックする．

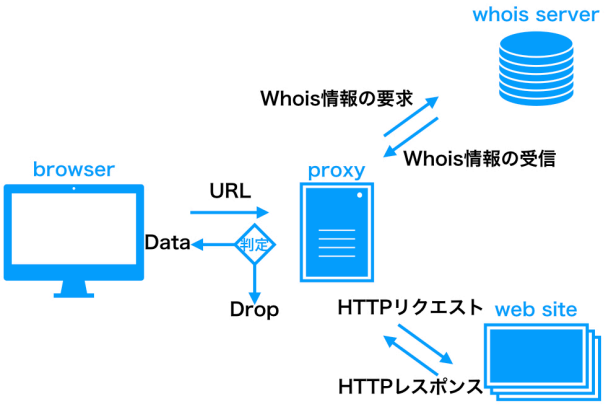

図3 システム構成図

本システムでサイトが安全か危険かを判断する式を以下に示す．先行研究[3]を参考にし，表1のi番目の特徴量に与えた重みw_iと，特徴量をカテゴリ化して割り当てた定数d_iを掛けた合計値が閾値θを超えた時にサイトを安全と判断する．

$$\sum_{i=0}^{7} w_i \cdot d_i > \theta$$

ただし，ここでw_iは判定プログラムで交差検証を繰り返し，発見的手法で定めている．計200件中2割を学習データとし，残りの8割をテストデータとした時の結果を表6に示す．FNは正規サイトを危険と判断した件数，TNはフィッシングサイトを安全と判断した件数である．それぞれ80件中設定されている件数を示す．その結果によって求まった全体正解率Accuracyは95%であり，有用性があると言える．

表6 提案検出システムの精度

	1区間	2区間	3区間	4区間	5区間	平均
FN	8	7	7	2	5	5.8
TN	2	1	2	5	1	2.2

4　おわりに

フィッシングサイトのドメインとHTTPレスポンスヘッダに関するデータは正規サイトのデータと異なり，特有の傾向がある事が分かった．フィッシングサイトを作成するためのキット等に焦点を当て，その特徴を求める事でより検出精度を向上する事を今後の課題とする．

参考文献

[1] フィッシング対策協議会，"フィッシングレポート2018"(https://www.antiphishing.jp/report/pdf/phishing_report_2018.pdf，2018年12月参照)

[2] phishtank(https://www.phishtank.com)

[3] 中村元彦,寺田真敏,千葉雄司,土井範久,“プロキシを利用した HTTPリクエスト解析によるフィッシングサイト検出システムの提案“ 情報処理学会論文誌 Vol.48 No.10,2007.

[4] MDN web docs HTTPヘッダー(https://developer.mozilla.org/ja/docs/Web/HTTP/Headers, 2018年12月参照)