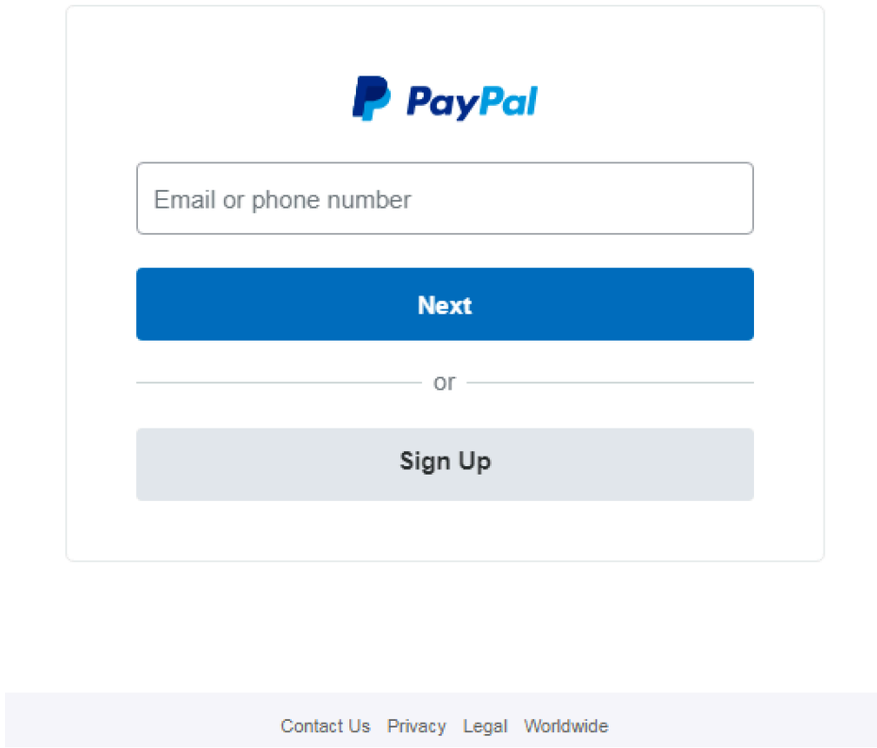
ドメイン情報とHTTPレスポンスヘッダに基づくフィッシングサイトの識別と評価

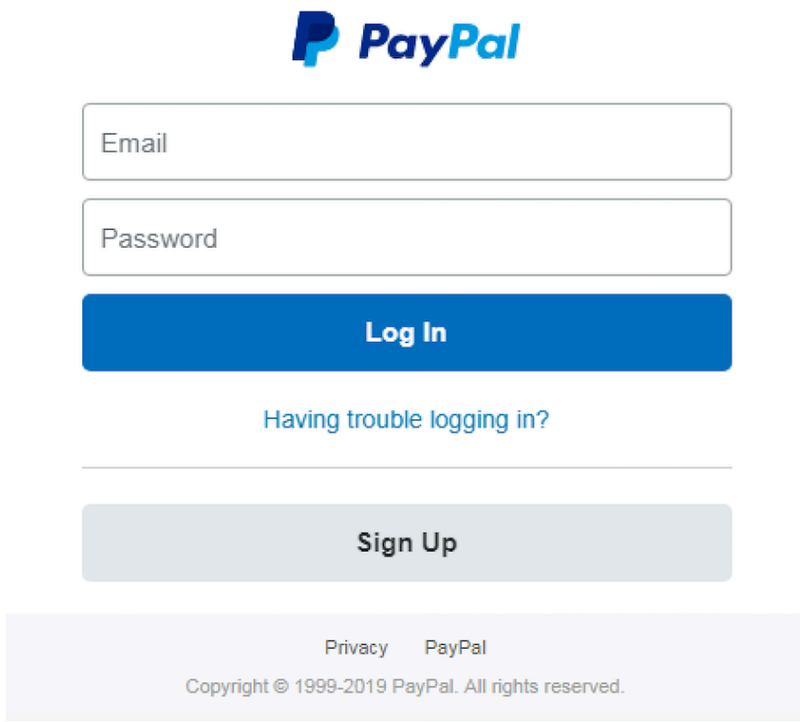
桜井啓多

正規サイトとフィッシングサイト

正規サイト



フィッシングサイト



サイトの生存期間: 7134日間

x-xss-protection: 設定あり

x-frame-options: 設定あり

サイトの生存期間: 189日間

x-xss-protection: 設定無し

x-frame-options: 設定無し

研究背景

- フィッシングとは,攻撃者がメール等で本物のサイトと同じ様な偽のWebサイトへのリンクを送信し,騙されたユーザがそのサイトに入力してしまった情報を盗む行為である.

- フィッシング対策協議会のフィッシングレポート2018(https://www.antiphishing.jp/report/pdf/phishing_report_2018.pdf)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URL件数は約5000件に上る.

(https://www.antiphishing.jp)

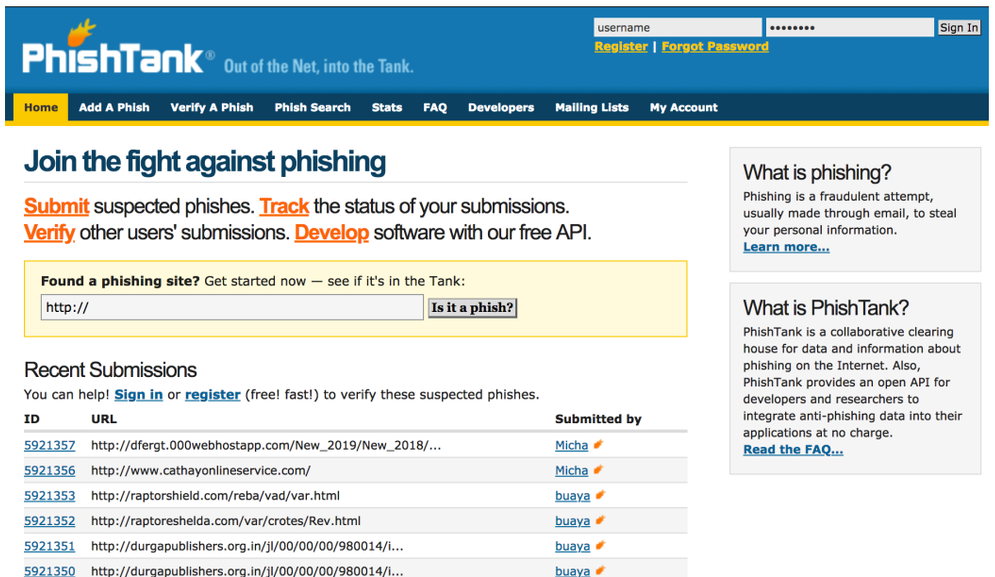
研究方法

- フィッシングサイトの検出を支援するため,対象のサイトが安全か否かを判定するプログラムを作成した.
- 判定を行うための指標として,フィッシングサイトのドメインをphishtankより取得し,そのドメインの情報を100件分と企業のホームページなどの正規サイト100件の計200件のデータを収集した.

フィッシングサイト	100件
正規サイト	100件

先行研究: “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“

(“https://www.phishtank.com”)

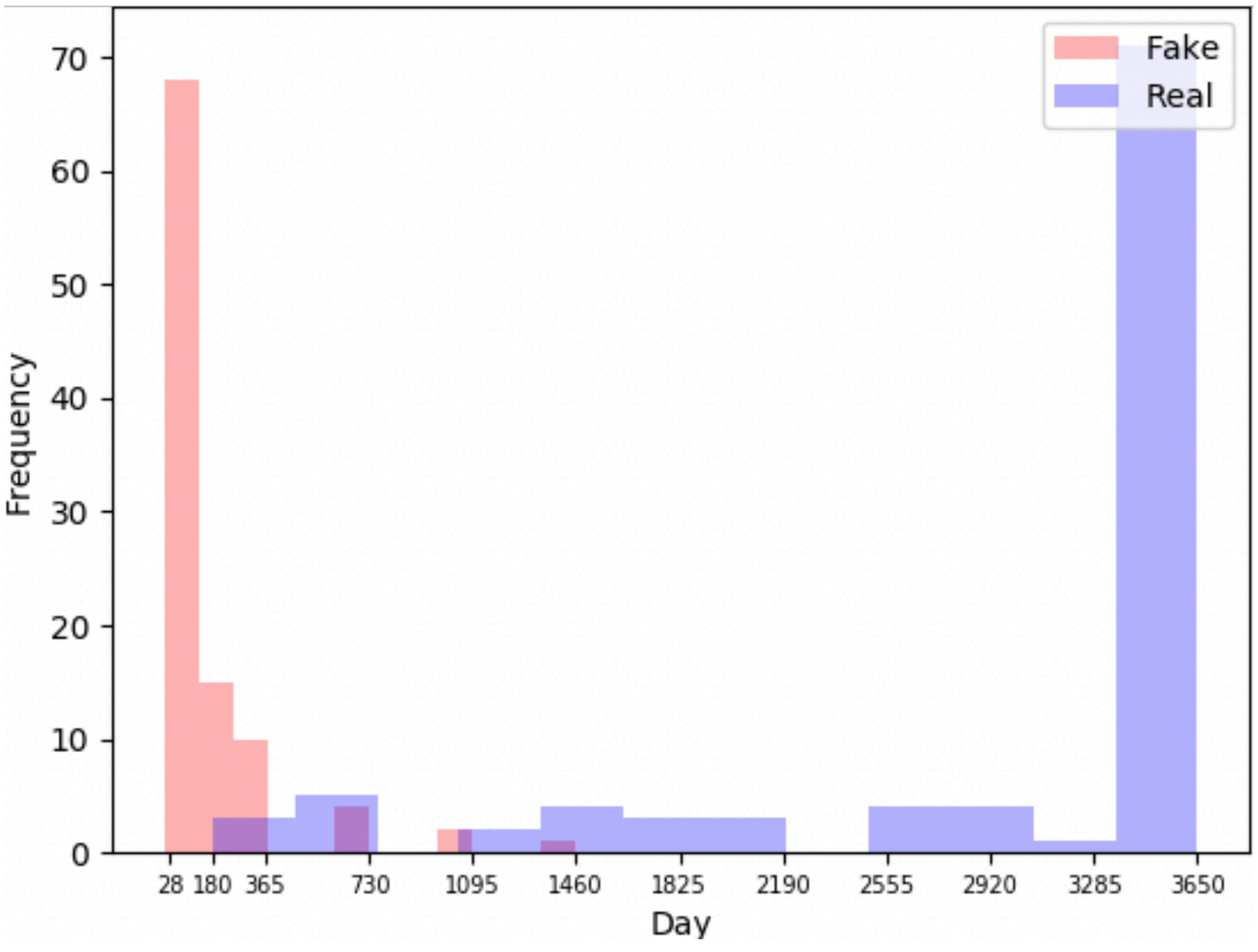


収集した特徴量の概要

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

ドメインに関するデータ①

”domain interval”: ドメイン作成日 - アクセス日

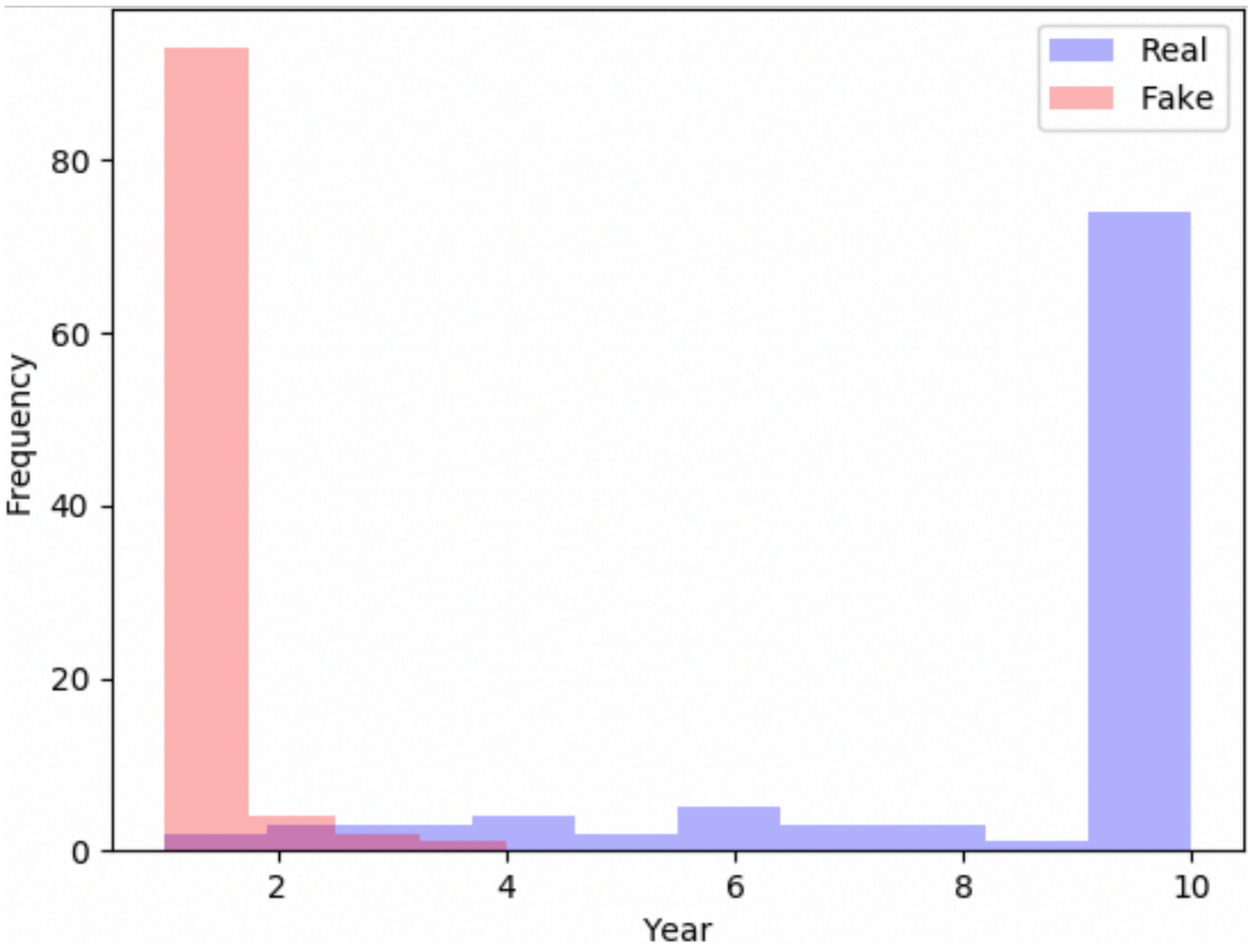


	正規サイト[日]	フィッシングサイト[日]
平均値	3089	135
中央値	3650	14
最頻値	3650	7

正規サイトは長く,フィッシングサイトは短い

ドメインに関するデータ②

”domain lifetime”: ドメイン期限日 - ドメイン作成日



	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

フィッシングサイトは短い運用
正規サイトは長い運用

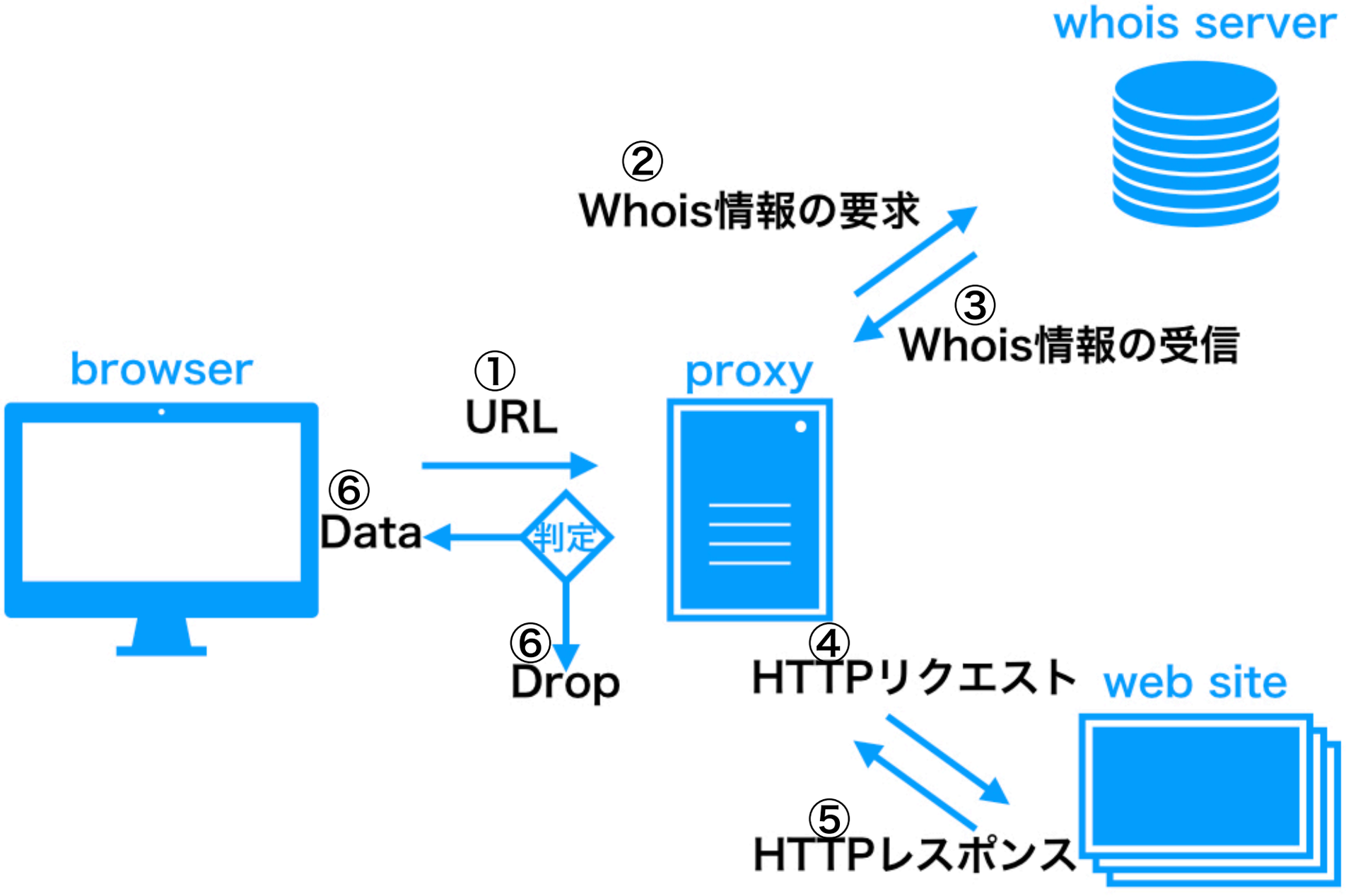
HTTPレスポンスヘッダに関するデータ

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

国に関するデータ

国名	正規サイト	フィッシングサイト
アメリカ	56	40
日本	40	2
アイルランド	2	0
台湾	0	20
オランダ	0	6
イギリス	0	4
パナマ	0	4
ロシア	0	4
カナダ	0	3
韓国	0	2
モロッコ	0	2
フランス	0	2
その他(1件のみ)	2	9
不明	0	2

システム概要


whois server
②
Whois情報の要求
③
Whois情報の受信
browser
①
URL
proxy
⑥
Data
判定
⑥
Drop
④
HTTPリクエスト
web site
⑤
HTTPレスポンス

判定方法

安全か判定する式

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

wi: 特徴量に与えた重み

di: カテゴリ化して割り当てた定数

θ: 閾値

(例) domain_intervalの定数di

期間	定数
1週間以下	0
1ヵ月以下	10
2年以下	50
4年以下	70
6年以下	80
7年以上	100

判定結果

FN	5.8
FP	2.2

- FN: 正規サイトをフィッシングサイトと判断した件数
- FP: フィッシングサイトを正規サイトと判断した件数

$$Accuracy: \frac{TP + TN}{TP + TN + FP + FN}$$

全体正答率Accuracy: 95%

まとめ

- 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.

- ドメイン,HTTPレスポンスヘッダ,IPアドレスの割り当て国を用いた識別は正答率95%となり,有用性があった.

- 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.