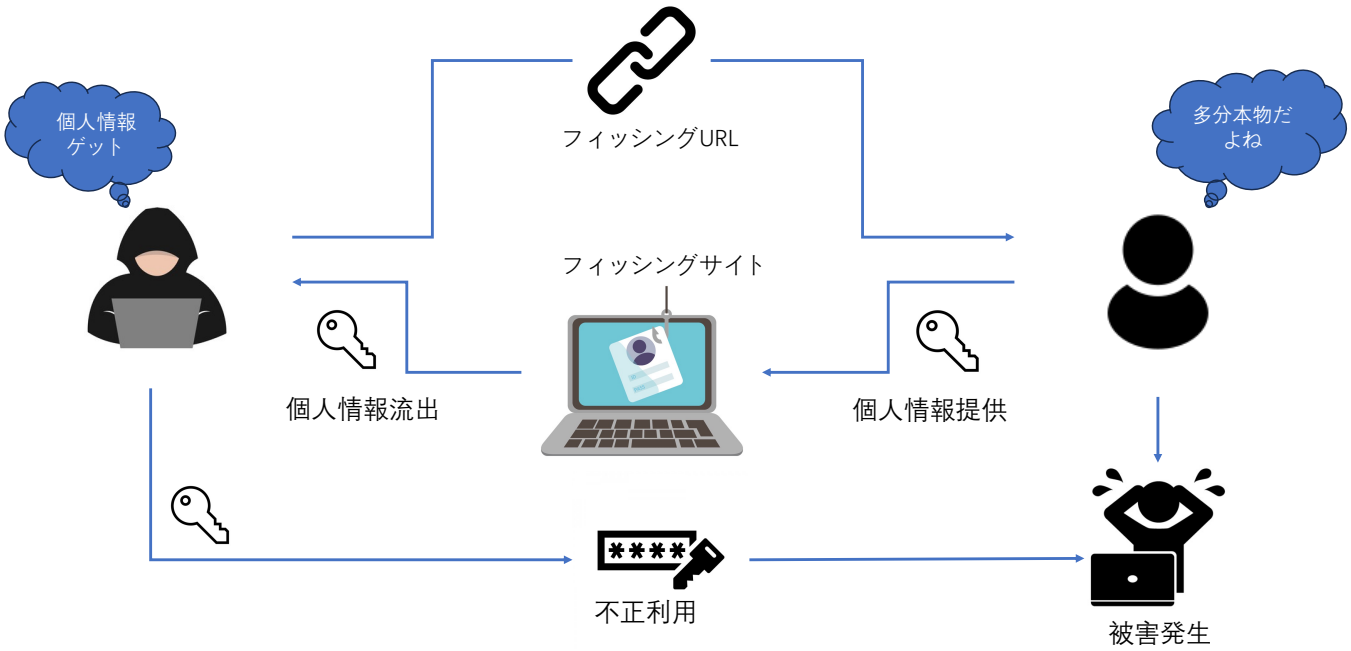
フィッシング検出方式の提案と JPCERT/CC フィッシングデータセットを用いた評価

明治大学　総合数理学部

菊池研　4年　YANG LIYI

研究背景


個人情報
ゲット
フィッシングURL
多分本物だ
よね
フィッシングサイト
個人情報流出
個人情報提供
****
不正利用
被害発生


au
17:19
100%
connect.auone.jp
au ID
ログイン
au ID(電話番号/任意のID/メールア
パスワード(8桁以上英数記号)
パスワードを表示する
ログインしたままにする
ログイン
au ID・パスワードを忘れた
au ID新規登録


au
15:47
100%
myau-net.com
au ID
ログイン
au ID(電話番号/任意のID/メールア
パスワード(8桁以上英数記号)
パスワードを表示する
ログインしたままにする
ログイン
au ID・パスワードを忘れた
au ID新規登録

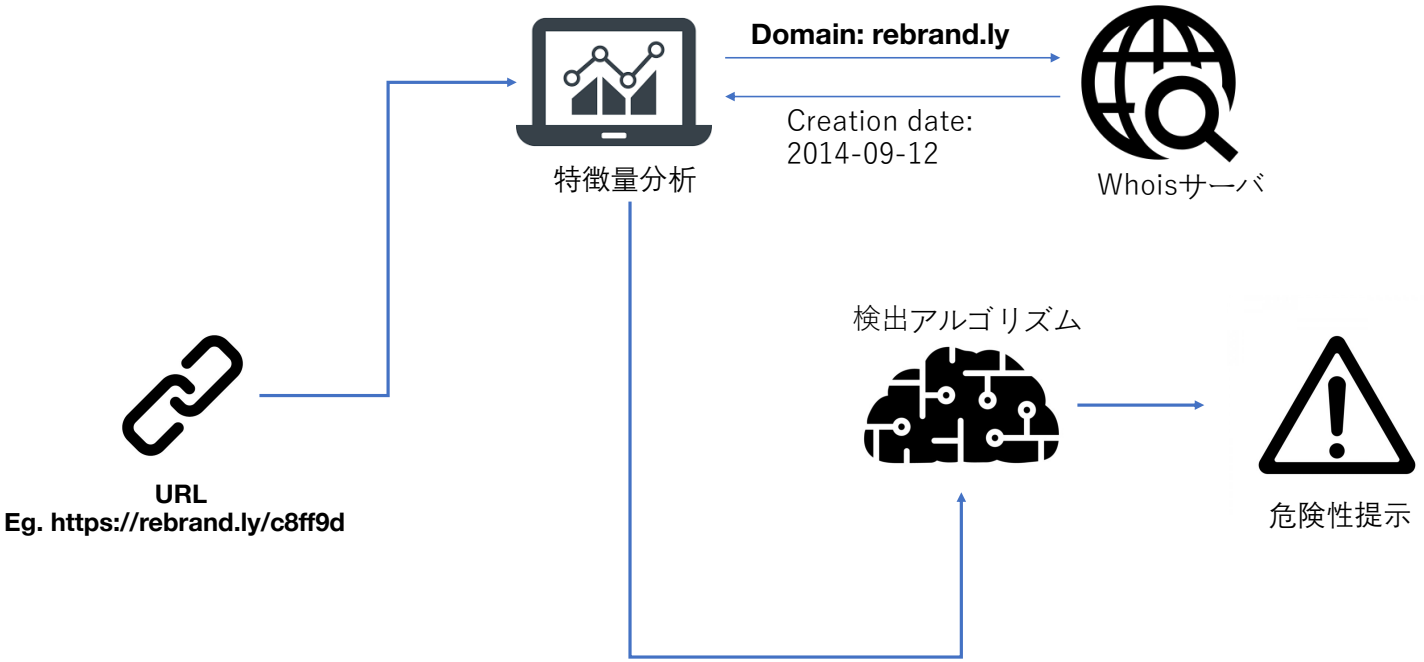
先行研究と問題点

- 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, pp3365-3374, 2007.
  - 特徴：**既存サービスNetcraft を利用**
  - 問題点：Netcraftは外国のサービスであり、正しく日本のフィッシングURLを検出できるかはまだ検証されてない

- 桜井啓多，“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”，2018年度菊池研究室卒業論文，2018.
  - 特徴：**HTTPレスポンスヘッダを利用**
  - 問題点：URLのドメイン部分だけに集中

提案方式

- 日本のフィッシングURLに特化したフィッシング検出システム



SSL risk	Random risk	Keyword risk	TLD count	Splited length	Whitelist risk	Domain duration	Similarity risk	HTTP status
0	2	2	1	2	0	3362	0	2

判断条件（点数付け）

- SVM（40点/0点）
  - **SSLリスク**
  - **ランダムリスク**
  - **キーワードリスク**
  - **TLDカウント**
  - **Splited length**
  - **ホワイトリストリスク**
- Domain duration（25点/15点/0点）
- Similarity risk（10点/0点）
- HTTP status（10点/0点）
- Total Point>55 ⟹ 危険

使うデータセットとやること

- データセット
  - 海外のフィッシングURL：Phishtank
  - 日本のフィッシングURL：JPCERT
  - 安全なURL：Kaggle
  - ホワイトリスト：Tranco
- やること
  - **既存サービスNetcraftの調査**
  - 海外のフィッシングURLと日本のフィッシングURLの違いを明らかにする
  - **新しい特徴量の提案**
  - 日本のフィッシングURLに特化したフィッシング検出システムの開発

調査結果

- Netcraft（1200個ずつ）

	警告件数	割合
JPCERT	141	11.8
Phishtank	676	56.3

警告確認

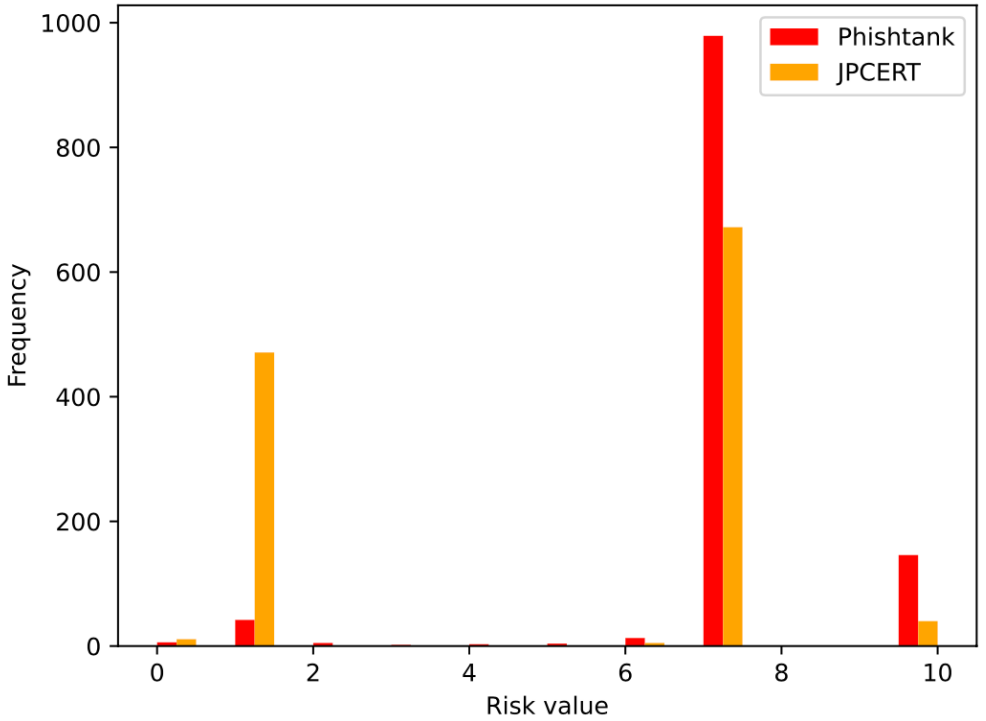
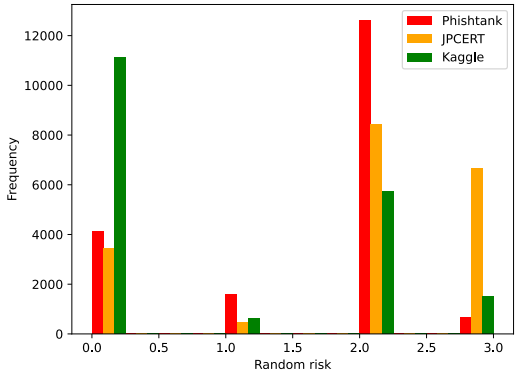
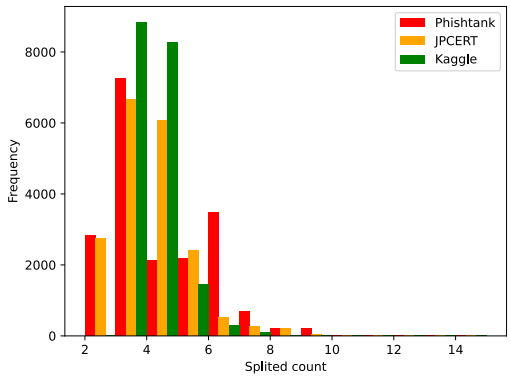


図 4.1　Netcraft のリスク値の分布

特徴量

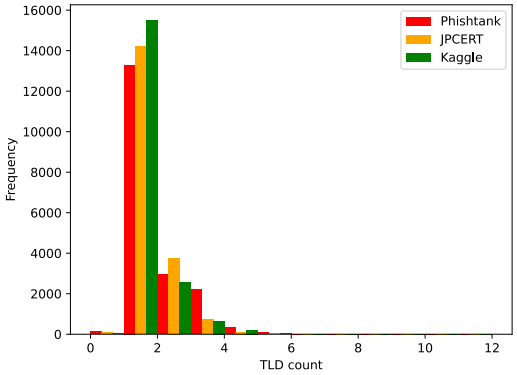
	SSLハイリスク割合
JPCERT	12.9
Phishtank	8.2
Kaggle	0

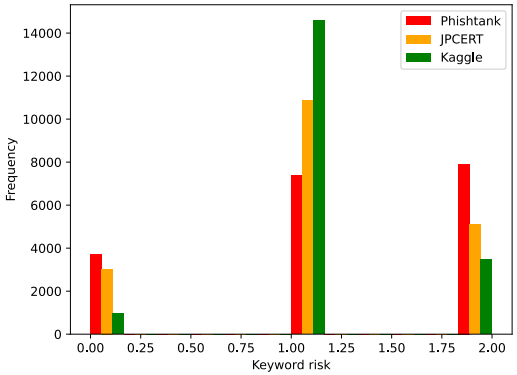




GOOD

BAD





	Whitelistハイリスク割合
JPCERT	97.7
Phishtank	77.8
Kaggle	96.1

評価

- JPCERT, Phishtank, Kaggleデータセットから URLをランダムに500個ずつサンプリングして用いて, 5のクロスバリデーションを行った。

	JPCERT	Phishtank
Precision	0.79	0.62
Recall	0.83	0.35
F score	0.81	0.45
Accuracy	0.81	0.57

結論と考察

- Netcraftは日本のフィッシングサイトに弱い
- **フィッシング対策は一つのサービスに頼らず, 地域に対応した複数のサービスを運用すべき**
- 今後, URLベースだけでではなく, 多様な方式で適切な特徴量を定め, より質の高いシステムを構築していきたい

- SSLリスク（良い特徴量）

表 3.1　SSL risk の例

URL	先頭部分	SSL risk
http://lphvnlopuh.duckdns.org/	http	1
https://lphvnlopuh.duckdns.org/	https	0

表 4.2　SSL リスク

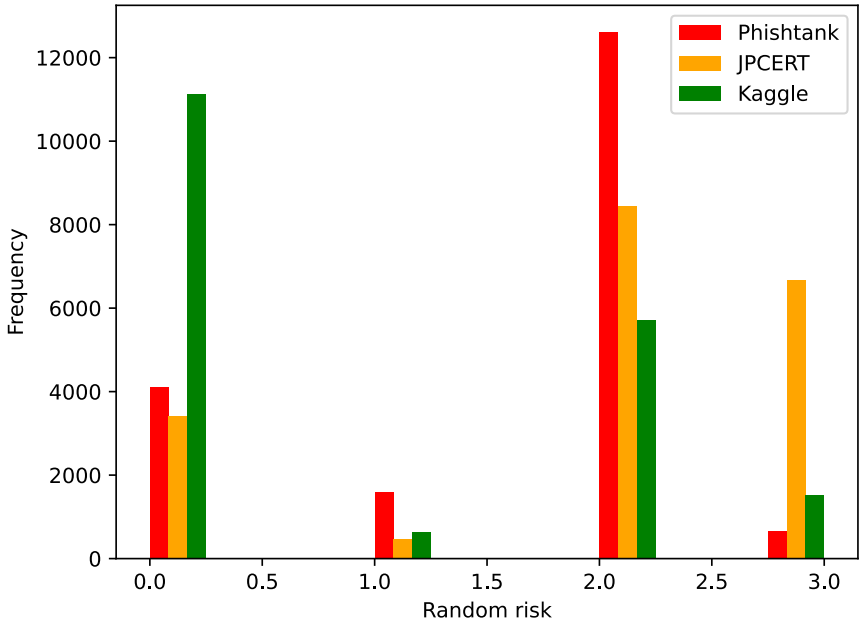
	positive	%
JPCERT	2442	12.9
Phishtank	1551	8.2
Kaggle	0	0

- ランダムリスク（良い特徴量）

$$Randomrisk = \begin{cases} 2 & \text{if } \min(P_{URL}) < 0.015, \\ 1 & \text{if } 0.015 \leq \min(P_{URL}) < 0.025, \\ 0 & \text{if } \min(P_{URL}) \geq 0.025. \end{cases}$$

表 3.2　Random risk の例

URL	評価最小値	Random risk
https://wzjdayup.xyz	0.0049	2
https://wwwepns.tanhehe.com/jp.php	0.0235	1
https://www.twitter.com	0.0260	0



- キーワードリスク（悪い特徴量）

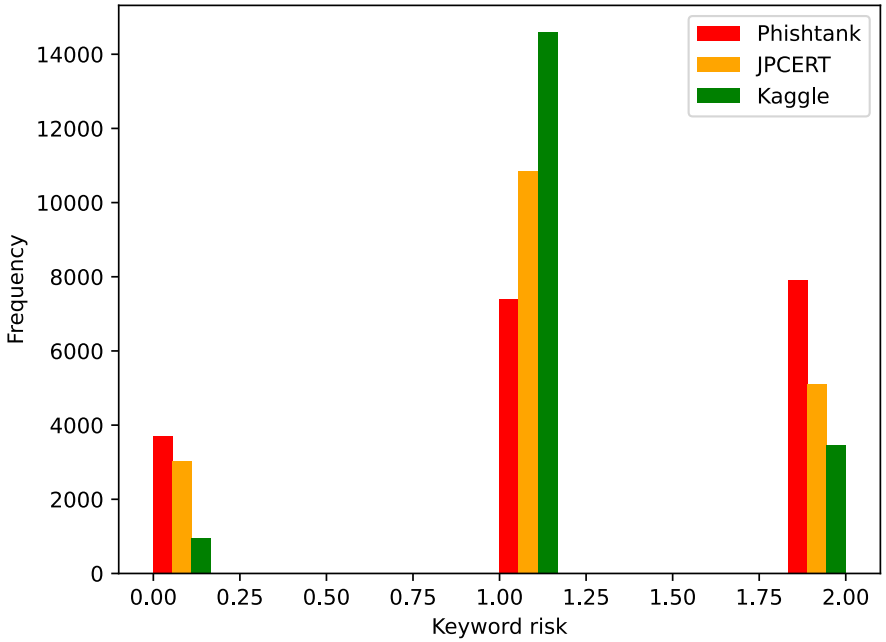
表 3.3　キーワードリストの一部

Description	Keyword
エポスカード	eposcard
楽天	rakuten
ログイン	login

$$Keywordrisk = \begin{cases} 2 & \text{if } \max(Q_{URL}) > 85, \\ 1 & \text{if } 70 \leq \max(Q_{URL}) \leq 85, \\ 0 & \text{if } \max(Q_{URL}) < 70. \end{cases}$$

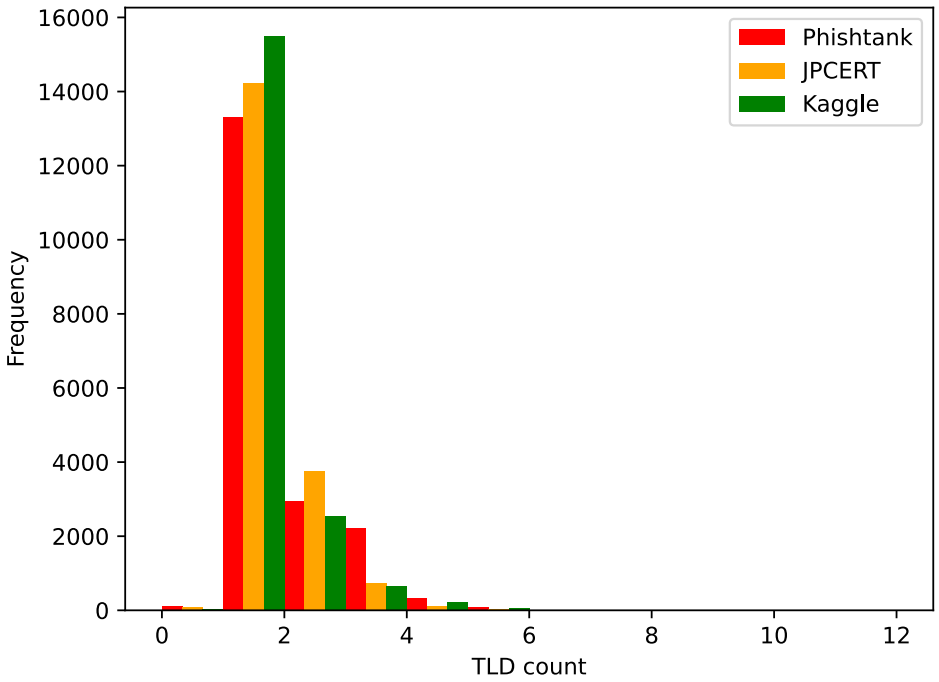
表 3.4　Keyword risk の例

URL	評価最大値	Keyword risk
https://www-cr-mufg-jp.kia8k.com/mufgcard/newsplus/	100.0	2
https://www.tmall.com	72.0	1
https://qwepo.xyz/	67.5	0



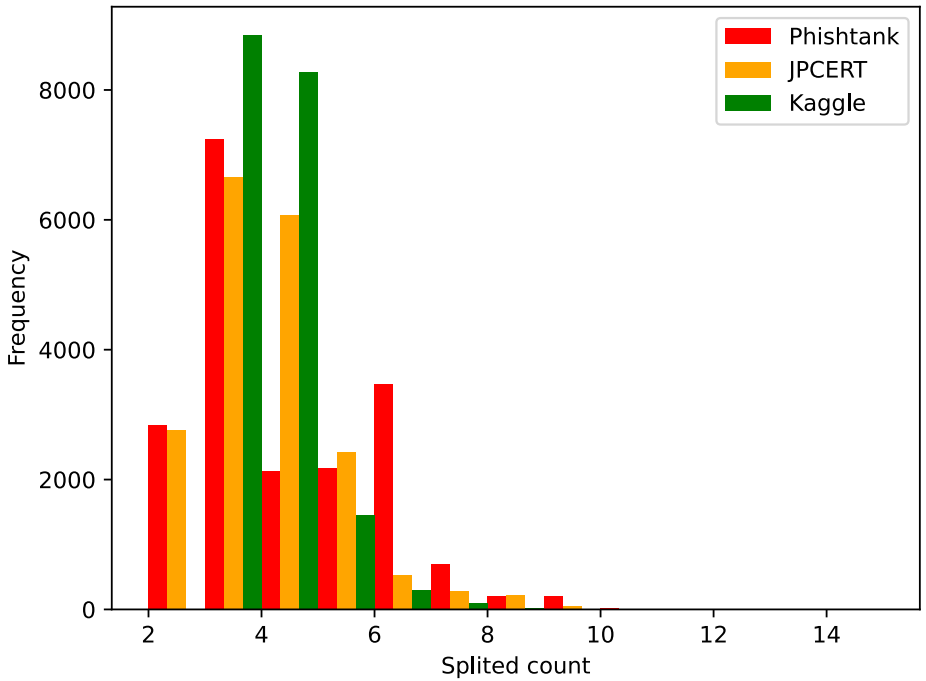
- TLDカウント（悪い特徴量）

例えば,
"https://newplenty.com.cn/jp" の場合はカウントされる部分文字列はcom, cn, jpであり, TLD countは3である



- Splited length（良い特徴量）

例えば, "https://info-e-orico.nftsgiant.com/" のドメインはinfo, e, orico, nftsgiant, comに分割され, Splited lengthは5である



- ホワイトリストリスク（悪い特徴量）

表 3.5　Whitelist risk の例

URL	ドメイン	Whitelist risk
http://lphvnlopuh.duckdns.org/	lphvnlopuh.duckdns.org	1
https://www.twitter.com/	twitter.com	0

表 4.3　Whitelist リスク

	positive	%
JPCERT	18564	97.7
Phishtank	14790	77.8
Kaggle	18256	96.1