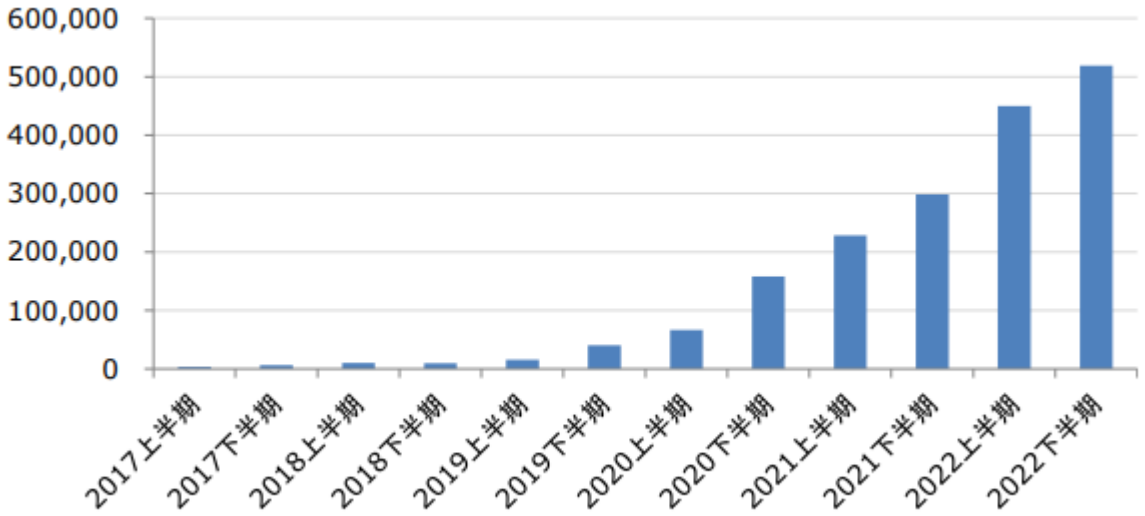
Phishing URL 攻撃パターンの自動分類とその評価

山本 悠太　菊池浩明
明治大学　総合数理学部

研究背景

- 近年フィッシング攻撃による被害は増加傾向で2022年度は968,832件のフィッシングサイトが報告されており, これは前年度の約2倍.



フィッシングサイト報告事例数

フィッシングサイトの例

研究目的

- URL生成パターンを自動分類し, 攻撃者を読み解き, 流行している攻撃手法を把握すること.
- 日本と海外のPhishing URL傾向の差を明らかにすること.

Phishing URLパターン例

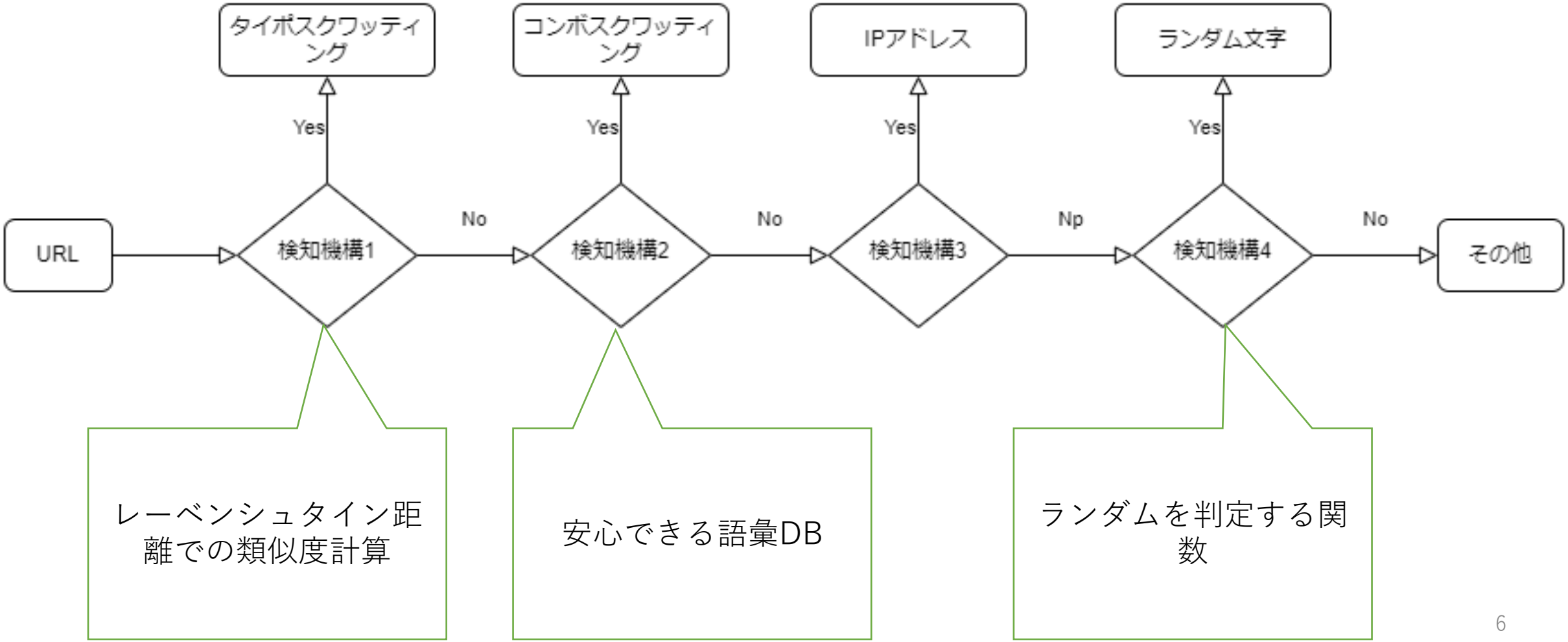
	例	オリジナルのURL	定義
タイポスクワッティング(Typo-squatting)	go**g**gle.com	Google	標的URLの打ち間違い(Typo)の意
コンボスクワッティング(Combo-squatting)	google-**secure**-**support**.com	Google	標的URLに安心できる単語を組み合わせる(Combo)の意
IPアドレス	192.168.0.1	N/A	IPアドレスの直打ち
ランダム文字	**hdashkjafdgs**.com	N/A	ランダム文字でパディング

困難点

- 打ち間違いを機械的にどう判断するか？
  - 例)amazom, amazoon, amazn

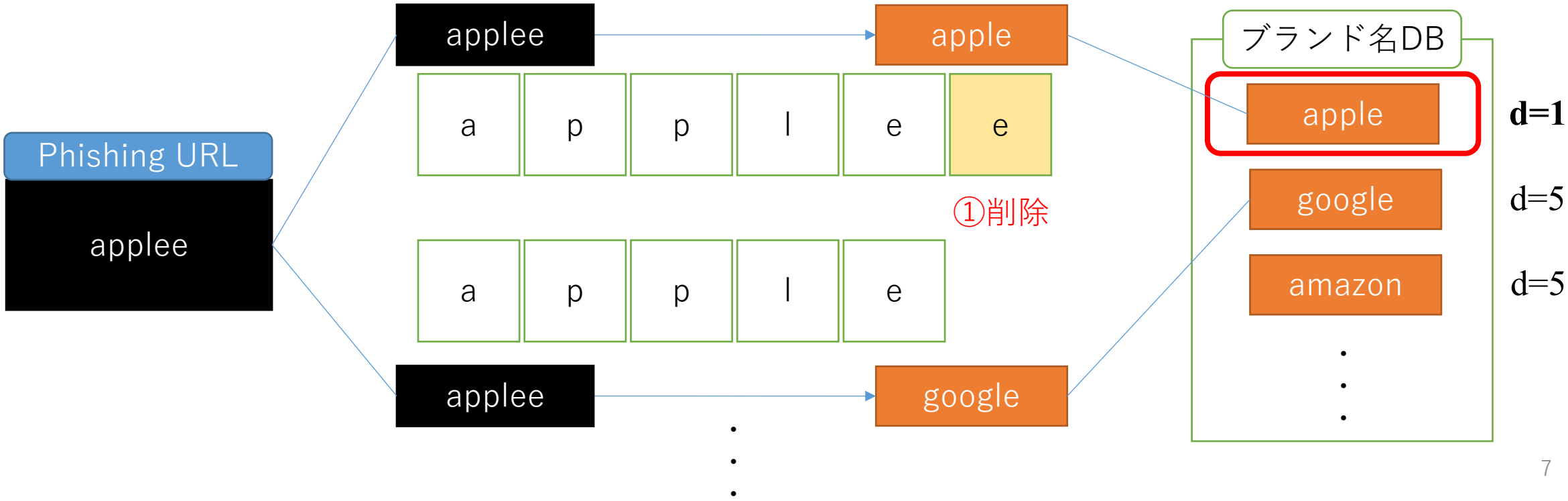
- Combo-squattingをどう判断するか？
  - 「標的URL」に「安全である単語」を付加して生成した、とどう判断する？

解決方法


タイポスクワッティング
コンボスクワッティング
IPアドレス
ランダム文字
Yes
Yes
Yes
Yes
URL
検知機構1
No
検知機構2
No
検知機構3
Np
検知機構4
No
その他
レーベンシュタイン距離での類似度計算
安心できる語彙DB
ランダムを判定する関数

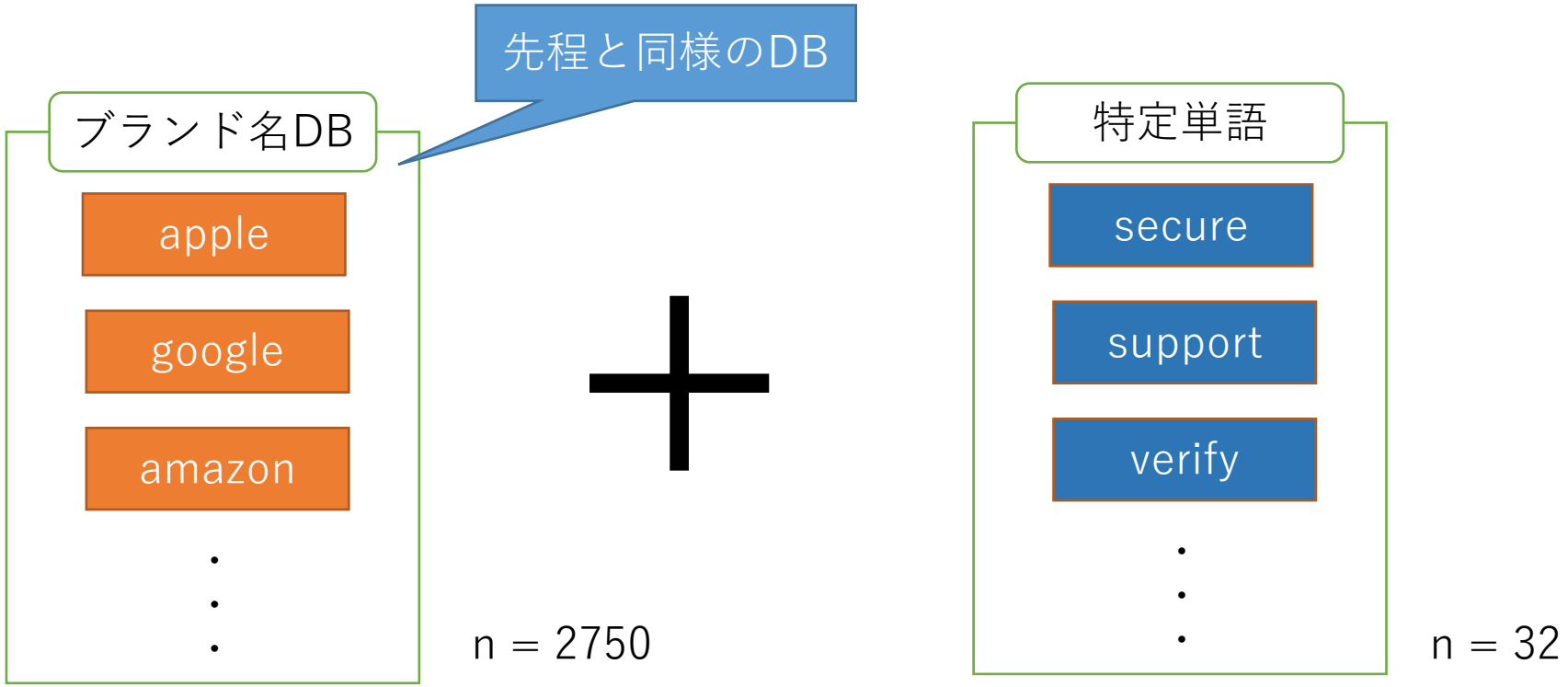
検知機構1:タイポスクワッティング検知

- レーベンシュタイン距離を使用する
- 比較対象となるブランド名DB
  - 「偽装元のブランド名」で検索したときの最上位のサイトのドメイン名をスクレイピングによって収集



検知機構2:コンボスクワッティング検知

- コンボスクワッティングは, 正規ブランド名＋特定単語が成立していた場合検知する.



検知機構3, 4:IPアドレス検知、ランダム検知

- IPアドレス挿入　例)192.168.0.1
- ランダム文字は, URL文字列の文字遷移確率pを計算し, p>0.075未満であればランダムとみなす.
- その他

評価実験

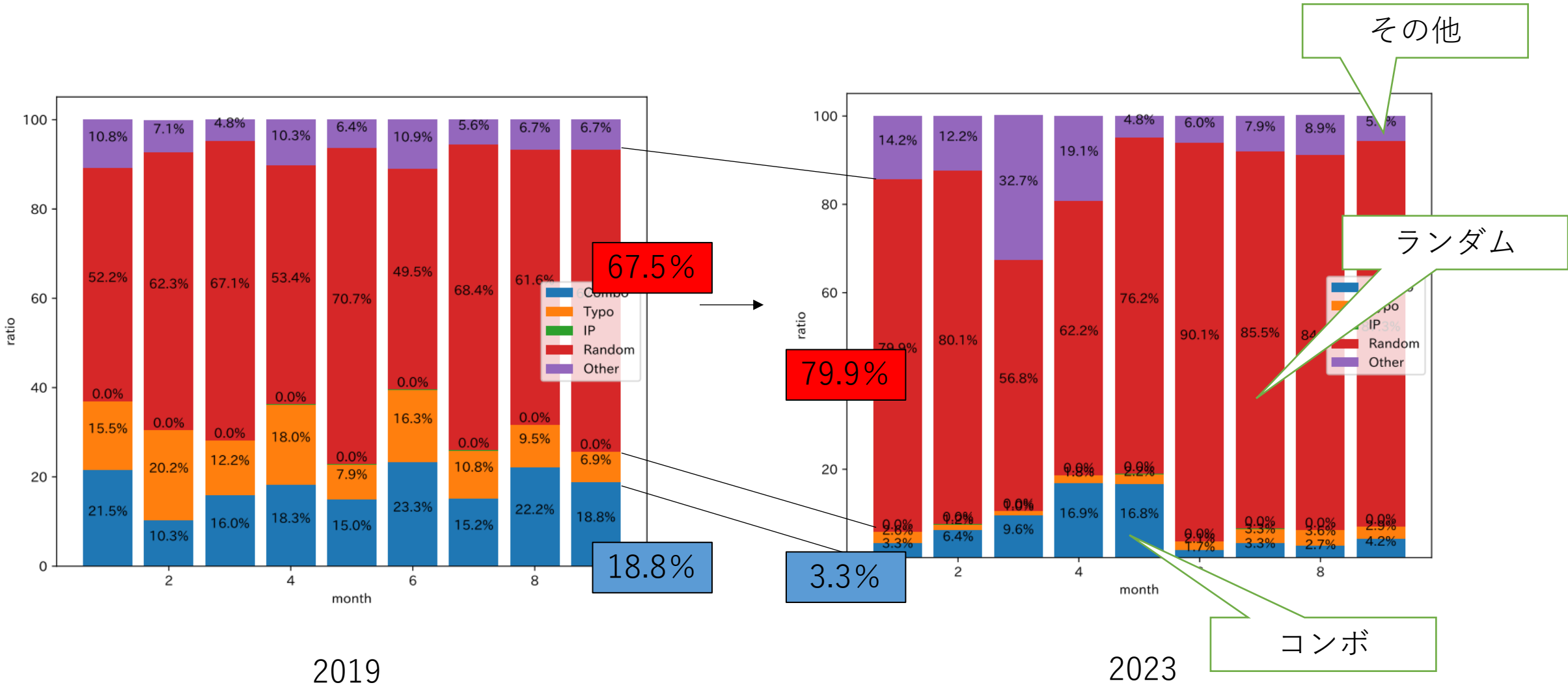
実験1:国内の生成パターン割合の変化を調査

	データセット	期間
国内	JPCERT	2019年3月-2023年9月

実験2:国内外の生成パターン傾向の比較

	データセット	レコード数	期間
国内	JPCERT	6,024	2023年9月度
国外	PhishTank	38,962	2023年11月度

結果1　国内の生成パターン割合の変化



結果2　国内外の生成パターン比較

	Typo	Combo	IP	Random	Other
国内 JPCERT	**5.4**	6.9	0.0	**82.6**	5,0
国外 PhishTank	**24.8**	6.8	0.0	**59.3**	9.1

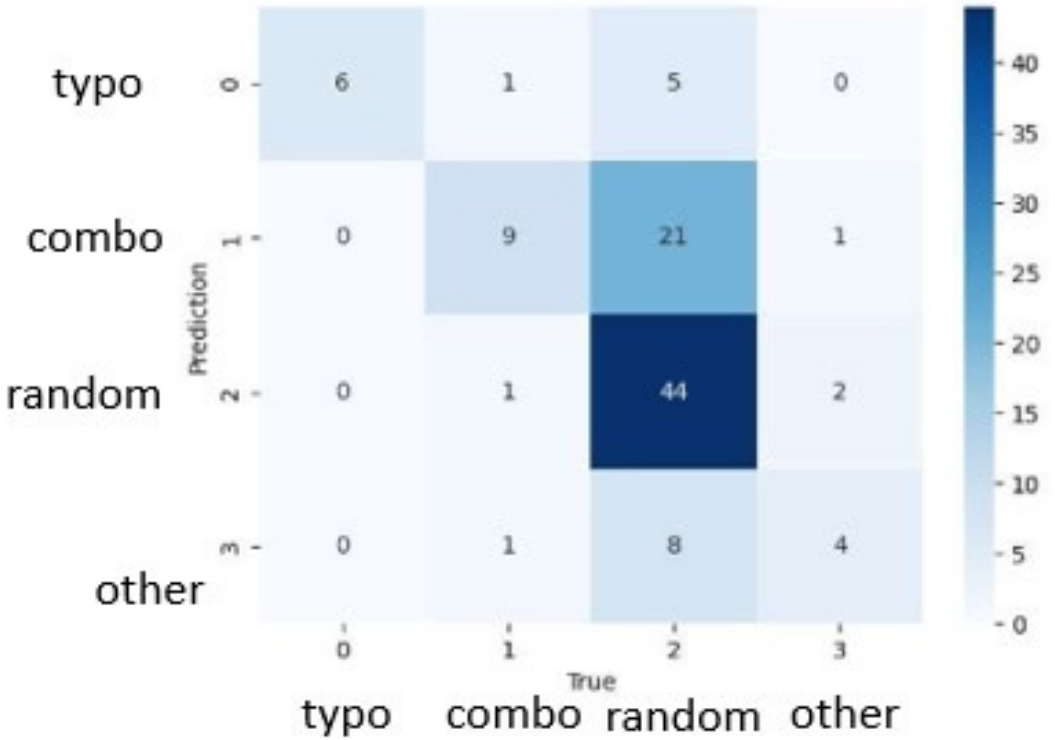
(単位:%)

	Typo	Combo	IP	Random	Other
国内 JPCERT	325	416	0	4976	301
国外 PhishTank	9663	2649	0	23104	3546

(単位:件)

検知機構の精度調査

- ランダムサンプリングしたPhishing URL102件について、目視での検知機構の精度調査を行った.
- acc = 0.61



結論

- 本研究ではPhishing URL攻撃手法パターンの自動判別手法を提案した. 国内での生成手法は, ランダム文字が20％も増加していた. また, 海外の攻撃パターンは日本と比べて20％程度タイポスクワッティングが多かった.

- 今後の課題: 検知機構を並列構造へと再構築すること, 検知機構の精度向上.