偽サイト検出のためのWebクローリングシステムの開発

Development of a web crawling system for detecting fake sites

加藤一樹・法制倫理分科会・情報セキュリティ大学院大学

1．背景

近年、新型コロナウイルスの影響により、マスク等の衛生用品やゲーム機等の巣ごもり消費関連の特需から、金銭を騙し取ったり、個人情報を窃取する偽ショッピングサイトが急増している。

そのため、警察などの捜査機関でも日々対策を行っているが、偽ショッピングサイトを把握する契機は、被害者からの相談や捜査員等の目視によるものが大半となっており、把握するまでに多くの時間と労力がかかっているのが現状である。

2．目的

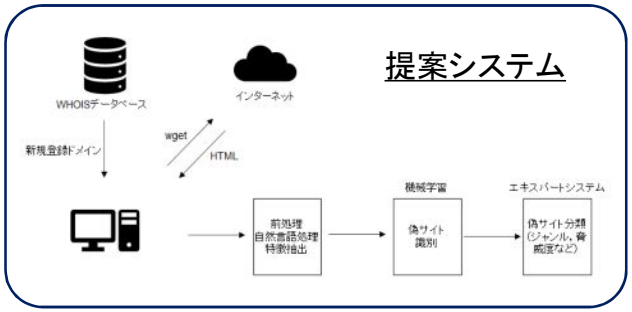
新規登録されたドメインをもとに、偽ショッピングサイトの検出を行い、プロセス全体を自動化することで、検査対象の広範化と省力化によるスループット向上を目指す。

具体的には、新規登録されたドメインから推定したWebサイトを偽ショッピングサイトの検査対象とし、機械学習等の人工知能技術を用いて識別させる。

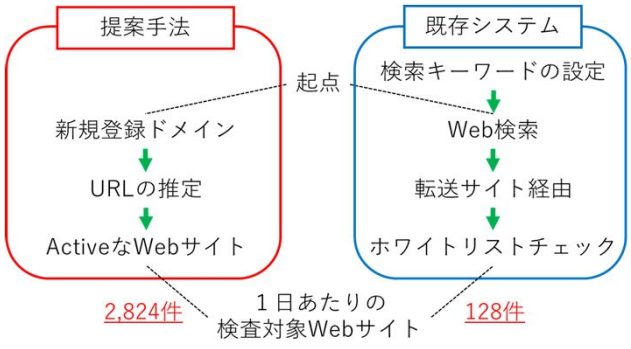
これにより、偽ショッピングサイトの早期発見が可能となり、被害者が出る前に対策を行うことやサイト作成者の検挙に向けた捜査時間を得ることができる。

3．提案手法

① 新規登録ドメインの取得
② URLの推定
③ HTMLデータの取得と前処理
④ 自然言語処理・特徴抽出
⑤ 機械学習による識別
⑥ エキスパートシステムによる分類



検査対象の広範化



4．まとめ

10日間の実験の結果、新規登録ドメイン（1,191,422件）の内、28,240件のWebサイトのHTMLデータを取得することができた。

これにより、既存のシステムと比較して、インターネット上のより多くのWeb サイトを検査対象として、偽ショッピングサイトを検出することが可能となったことから、新規登録ドメインから偽ショッピングサイトを検出する手法は有効であると言える。

しかし、検出された多くのWeb サイトを検査対象とするには、既存システムの処理速度では、日々の運用に耐えることは厳しいことから、分散処理や識別モデルの運用方法等の検討を行う必要がある。

5．課題

- 実利用に向けたUIと完全自動化
- システムの高速化
- エキスパートシステムによる詳細な分類
- ドメイン未取得サイトへの対応
- サブドメインやホスティングサービスへの対応