偽ショッピングサイト自動検出システムの開発と今後の研究

An automatic detection system for fake shopping sites and future research

堺　啓介・システム分科会・情報セキュリティ大学院大学

Abstract - The proliferation of fake shopping websites is on the rise, however law enforcement agencies, such as the police, have not been able to implement comprehensive solutions. Therefore, there is a need for efficient and early detection methods. In this research, we created an automated detection system for fake shopping websites using machine learning, starting with a list of newly registered domains. We were able to achieve early detection of fake shopping websites and increase the efficiency of detection efforts.

研究背景

偽ショッピングサイトは増加傾向で、昨今のニュースでもシーズンに応じた注意喚起がなされるなど、その脅威は猛威を奮っている。

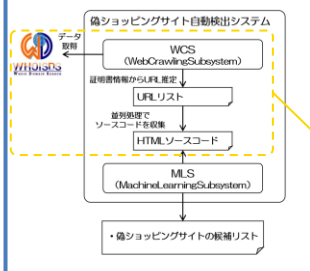
関連研究と本研究

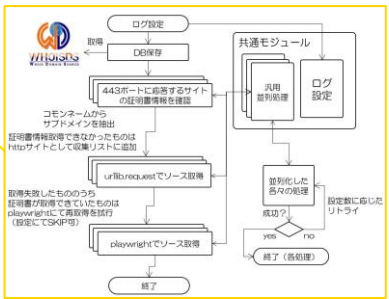
	偽サイト識別	WebCrawling
関連研究	・識別手法の研究は多数 ・データ収集はスコープ外	・長年に渡って、研究事例は多数 ・マシンスペックを考慮した研究例なし
本研究	低スペック環境*1でも効率的に動作するWebCrawlingと高速かつ高精度なMachineLearning*2を組み合わせた偽ショッピングサイト自動検出システム*3を実社会に適用	
目的	偽ショッピングサイト脅威の低減・無効化 具体的にはサイバーパトロールの業務効率化を目指す	

※1：中小規模な研究機関や官公庁では、何十台ものマシンで並列処理を行う環境の用意が困難
※2：IISEC（栗原氏）から研究を引き継いだJC3で作成したプログラムを改修して実験
※3：新規登録ドメインに着目したシステム構想は、IISEC（加藤氏）による先行研究による提案

システム設計

WCSが並列処理でデータを収集、MLSが偽ショッピングサイトの判定*5

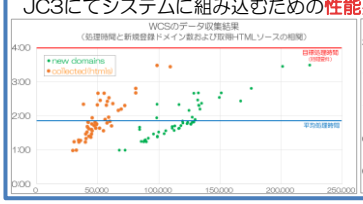


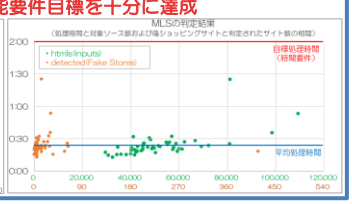


※5：WebCrawlingSubsystemの略でWCS、MachineLearningSubsystemの略でMLSと呼称する。
MLSの設計はJC3にて行っているため、詳細を省略する。

結果

JC3にてシステムに組み込むための**性能要件目標を十分に達成**





評価

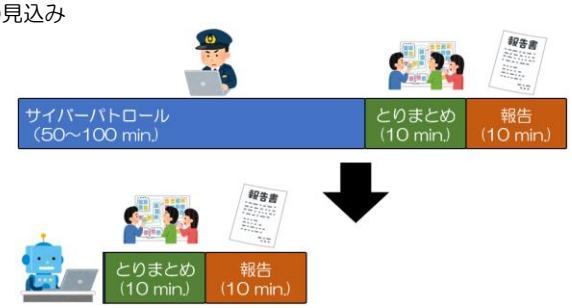
WCSの性能評価

スペック等を考慮して超概算で試算すると、
速度は**先行研究の約500倍**・**既存研究の約24倍**と推察され
データ取得率も**先行研究の約18倍**・**既存研究の約2倍**と大きく改善

項目	先行研究	本研究	Zowallaらの研究
仮想マシン台数	1	1	22
CPU	Intel Corei9-10900	Intel Xeon E-2324G	Intel Xeon E5-2689
CPUコア数	10	2	8
メモリ	16 GB	32 GB	256 GB
データ取得速度	0.03 html/sec	7.58 html/sec	7~10 html/sec
データ取得率	2.37%	43.51%	19.76%

システム全体の評価

サイバーパトロール業務において最も時間を要していた、クローリングからWebサイトの判定を**一気通貫で自動化**することに成功し
・**約2~4人日/月の工数削減**・**約3.5倍~13倍の検出効率向上**
の見込み



項目	学生ボランティア	有識者	本システム
サイバーパトロール	100 min.	50 min.	0 min.
とりまとめ、報告	20 min.	20 min.	20 min.
検出数（FakeStores）	10	20	21
工数	5.00 人日	2.92 人日	0.83 人日
検出効率（FakeStores/min.）	0.08	0.29	1.05

まとめ

成果　**JC3の偽ショッピングサイト対策の取り組みとして採用されシステムに組み込まれて稼働中**

今後
・**産官学連携**プロジェクトとして共同研究を計画
・システムの改善、他の悪性サイトへの横展開

今後の研究案

・URL推定の改善（サブドメイン：証明書、コンテンツ：機械学習）
・TwitterやYahooのトレンドとGoogle Custom Search API等を利用した新規登録ドメイン以外の入力ソースの検討
・検出した偽サイトのライフサイクル分析、証明書情報等の分析
・証明書情報に着目した、偽サイト識別用ブラウザアドオンの開発