fastTextとLightGBMを用いた偽ショッピングサイト自動検出システムの開発

堺 啓介[1]　竹重 耕介[2]　加藤 一樹[1]　栗原 直樹[3]　大野 克巳[2]　橋本 正樹[1,a)]

概要：近年，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，警察等の法執行機関では対処が必要であるが，現状では日々生成される偽ショッピングサイトの全体像や被害額等，実態の正確な把握すらできていない．本研究では，人工知能技術を用いた偽ショッピングサイト自動検出システムを開発し，偽ショッピングサイトのURL収集を効率化する．提案システムは，新規登録ドメインリストを起点に，検査対象とするウェブサイトのHTMLデータを収集した上で，機械学習により偽ショッピングサイトか否かを自動判定するものである．提案システムは，新規登録ドメインリストから1日平均約6万6千件の検査対象URLを推定し，平均約4万5千件のHTMLデータを収集することに成功した．また，同データについて，機械学習を用いて偽ショッピングサイトに該当するかを判別し，結果として，平均約20件の偽ショッピングサイトを自動的に検出することができた．

キーワード：偽サイト，機械学習，ドメイン，URL，Webクローリング

An Automatic Detection System for Fake Shopping Sites using fastText and LightGBM

Keisuke Sakai[1]　Kosuke Takeshige[2]　Kazuki Kato[1]　Naoki Kurihara[3]　Katsumi Ono[2]
Masaki Hashimoto[1,a)]

Abstract: In recent years, the number of fake shopping sites that scam people out of their money or steal their personal information has skyrocketed. However, the current situation is such that law enforcement agencies are unable to accurately grasp the overall picture of the fake shopping sites that are being created every day, including the amount of damage caused. In this study, we develop an automatic detection system for fake shopping sites using artificial intelligence technology to streamline the collection of fake shopping site URLs. The proposed system starts from a list of newly registered domains, collects HTML data of websites to be inspected, and automatically determines whether they are fake shopping sites or not through machine learning. The proposed system successfully estimated an average of 66,000 target URLs per day from the list of newly registered domains, and collected an average of 45,000 HTML data. The proposed system also uses machine learning to determine whether the data corresponds to a fake shopping site, and as a result, it was able to automatically detect an average of about 20 fake shopping sites.

Keywords: fake sites, machine learning, domains, URLs, web crawling

[1] 情報セキュリティ大学院大学
Graduate School of Information Security, Institute of Information Security
[2] 日本サイバー犯罪対策センター
Japan Cybercrime Control Center
[3] EY新日本有限責任監査法人
Ernst Young ShinNihon LLC
[a)] hashimoto@iisec.ac.jp

1. はじめに

近年，インターネットが普及し，サイバー空間上の犯罪が増加している．特に我が国においては，Web検索結果から，悪質なショッピングサイトへ誘導する手口や，正規のショッピングサイトを模倣した悪質なショッピングサイ

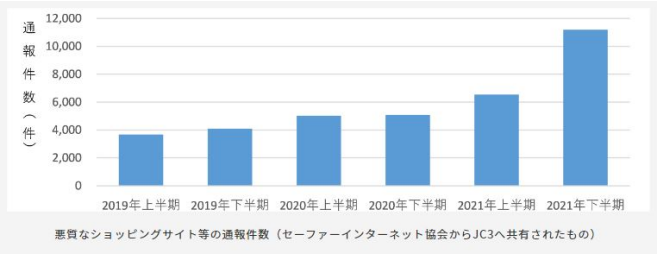

図 1 JC3 による悪質なショッピングサイト等に関する統計情報（2021 年）

Fig. 1 Malicious shopping site report by JC3(2021).

トによる被害が増加している．一般財団法人日本サイバー犯罪対策センター（Japan Cybercrime Control Center：以下，JC3）によると，2021 年に，セーファーインターネット協会 [1] から JC3 へ共有された悪質なショッピングサイト等の通報件数は，17,878 件となっており，2020 年の 10,095 件と比べ，7,783 件（約 77.1%）増加している（**図 1**)[2].

このような状況から，警察等の法執行機関でもその対処を試みているが，現状では，悪質なショッピングサイト等を把握する契機は，被害者からの相談やサイバーパトロール，すなわち捜査員等の目視によるものが大半となっており，結果として，偽ショッピングサイトの把握に多大なコストをかけているにも関わらず，ごく限られた成果しか出ていない状況である．

本研究は，日々の新規登録ドメインを起点として偽ショッピングサイトの検出プロセス全体を自動化することで，作業の省力化だけでなく，検査対象の広範化と偽ショッピングサイトの早期発見を可能とし，サイバー空間における偽ショッピングサイト脅威の低減・無効化に寄与することを目的とするものである．すなわち本研究においては，i) 日々新規に登録されるドメイン名を検査対象サイト収集の起点とすること，ii) 収集から判定に至る検査プロセス全体を自動化すること，により，インターネット上に新たに公開される Web サイトから自動的に偽ショッピングサイトの URL を早期に判別できるようになることを期待しており，結果として，被害者が出る前にセキュリティ事業者等に対する情報提供が可能となることを目指す．なお，本研究が対象とする偽ショッピングサイトは，偽サイトの一種と位置付けられる．

以降，本稿では，まず第 2 章において関連研究を説明する．次に第 3 章では，提案システムの設計と実装について述べ，その後の第 4 章ではその評価結果について説明する．最後に，本稿をまとめ，今後の課題について説明する．

2. 関連研究と本研究の位置付け

2.1 偽サイト識別手法に関する研究

近年，偽サイト検出に機械学習を応用する研究が非常に数多くなされている．これらの既存研究は，大きく i)URL・ドメイン名の文字列に着目するもの，ii)HTML のツリー構造に着目するもの，iii)HTML のソースコードに着目するもの，iv) その他の研究，と整理・分類することができる．

機械学習を用いた偽サイト検出手法は，一般的に Accuracy(検出の正確性）により評価されるが，昨今提案されている諸手法は Accuracy が 90%を超えるものがほとんどで，偽サイト検出問題に対する機械学習の有効性は実証済であると言える．

2.1.1 URL・ドメイン名の文字列に着目した研究

URL やドメイン名の文字列から偽サイトを分類・識別する研究は，近年非常に活発に行われており，数多くの試みがなされている．例えば，URL の文字列を特定のアルゴリズムで評価して偽サイトを識別する研究 [3]，Fast-flux, Double-flux, Domain-flux(DGA) 等利用した攻撃で観測されるドメインの特徴を機械学習させて偽サイト識別に役立てようという研究 [4]，ログインフォームが存在するページの URL 文字列に着目して機械学習を用いて偽サイトを識別する研究 [5] などがあり，その他にも非常に多くの報告がなされている [6]-[16]．先に述べた通り，いずれの提案においても，高い精度で偽サイトを検出できた旨が報告されている．

2.1.2 HTML のツリー構造に着目した研究

HTML のツリー構造から偽サイトを分類・識別する研究では，Alexa 順位をベースに上位サイトに限定して HTML ドキュメントオブジェクトモデル（DOM）の情報を収集して，機械学習で偽サイトを検出しようとした研究 [17], 楽天に関する 5 つのデータセット（サイズ：45,244,563,2033,3001）に対して SVM の機械学習モデルを適用して偽サイト識別を試みる研究 [18]，DOM のツリー構造を効率的に分析するためにカーネル法を適用した機械学習で偽サイトを識別しようとした研究 [19]，DOM の構造と CSS 構造を組み合わせて，機械学習で偽サイトを識別する研究 [20] などの事例がある．ツリー構造と機械学習による偽サイト識別についても，いずれの研究も高い精度での検出ができた旨，報告がなされている．

2.1.3 HTML のソースコードに着目した研究

HTML のソースコードから偽サイトを分類・識別する研究では，偽サイトのうち偽ショッピングサイトに特化して，機械学習を用いて偽ショッピングサイトを識別する研究 [21]，ソースコードのハイパーリンクに着目して，機械学習で偽サイトを識別する研究 [22]，COVID-19 に関して whoisds から 6321 の正規データ，domain tools dataset から同数の偽サイトデータを入手してデータセット化して Java Script や SourceCode，ページのコンテンツやスタイルに対して機械学習を試みて偽サイトを識別する研究 [23]，などの事例がある．

2.1.4 その他の研究

外部システム（検索エンジン等）を活用して偽サイトを識別する研究では受信メールと検索エンジンを組み合わせて偽サイトを識別する研究 [24] や Microsoft Reputation

Services (MRS) を活用して，URL を分類して偽サイトを識別する研究 [25] などの事例がある．

また，これまでに挙げた手法を複合させて偽サイトを識別する研究では，URL 情報とドメインの経過期間やサブドメイン，アンカー URL，IP アドレス等で偽サイトを識別しようとする研究 [26]，URL 情報とドメインに紐づけられた DNS に関連する情報に着目した研究 [27]，正しいサイト情報として Alexa 等の 6 サイトを，偽サイトとして Phishing Site URL 等の 6 サイトを参照し，計 3,980,870URL のデータセットを利用し，Linguistic Features, Human-Engineered Features, Deep-Web Features, URL Segmentation, Host-Based Features, Content-Based Features の観点を組み合わせて機械学習を試みた研究 [28] などの事例がある．

2.2 Web Crawling に関する研究

2.1 節で説明した偽サイト識別手法に関する様々な研究は，いずれも何がしかの研究用データセットを対象に識別精度の向上を試みるものであるが，実際に偽サイト検出システムを組み上げる場合には，検査対象とするデータを収集する機能が必要となる．本節ではこの観点から，Web Crawling に関する既存研究について説明する．

Olston らの研究 [29] では，Web Crawling を動作別に Incremental Crawling と Batch Crawling に大別し，商用 Crawler のほとんどを Incremental Crawling と分類した上で，Web Crawling の主な課題を網羅性と情報の鮮度にあると主張している．また，Castillo の研究 [30] では，既存の Web Crawling に関する研究が検索エンジンの内部機能として利用することを想定しており，研究用データ収集の観点からはあまり検討が進んでいない旨，問題提起をしている．Tchakounte らの研究 [31] は，検索クエリを起点に，URL の選択，URL の探索と重複の検出，ページのコンテンツの抽出と前処理，類似性の評価，保存等を行って偽サイト情報の収集を試みるたものであり，偽サイト用に設計された最初の Crawler であると主張している．

なお，Batsakis らの研究 [32] によれば，Web Crawler の重要な評価指標は Harvest rate（収集率）である．

2.3 関連研究のまとめと本研究の位置付け

本研究は，Web Crawler によるデータ収集と，HTML ソースコードに着目した機械学習により，偽ショッピングサイトの URL を検出するシステムを開発するものである．従って，偽サイト用の Web Crawler としては Batsakis らの研究に近く，検出手法としては，2.1.3 節で説明した HTML のソースコードに対する機械学習によるものに分類される．ソースコードに対する機械学習手法を選択した理由は，偽サイトのうちでも特に偽ショッピングサイトは，例えば，ID やパスワードを窃取する目的のフィッシングサイト（偽ログインサイト）とは異なり，URL やサイトのツリー構造ではなく，ソースコードそのものに着目しなければ判別が困難であるためである．他方で，ソースコードを収集するためには，Web サイトにアクセスして実際にレンダリングする必要があり，非常にコストのかかる作業であるため，偽ショッピングサイト自動検出システムの開発にあたっては Web Crawling に課題がある．

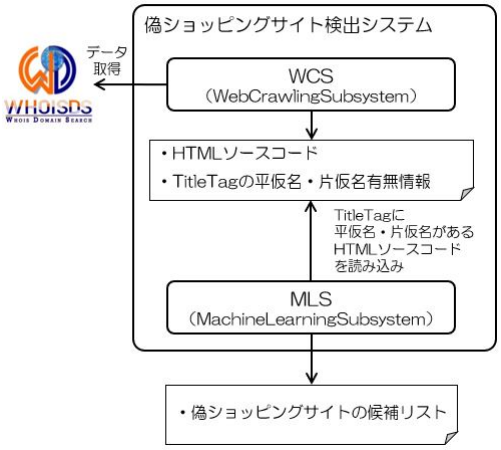

図 2 システムの全体概要
Fig. 2 System Overview.

また，先に述べた通り，既存研究の多くは予め準備されたデータセットを用いて機械学習による識別精度の向上を図るものであり，公開されている実際の Web サイトのデータを収集して識別するような実践的，一気通貫的なシステムは研究・検討が不十分であると考える．

そのため本研究は，以下に焦点を当てて検討を進めるものである．

(1) 偽サイト向けの効率的な Web Crawler を開発する．
(2) 機械学習による様々な識別手法を実際の Web サイトに対して適用し，偽ショッピングサイトの検出に適した手法を選択する．
(3) 1 と 2 を一つのシステムとして結合する．

3. 偽ショッピングサイト自動検出システムの設計

3.1 偽ショッピングサイト自動検出システムの概要

本システムは，WEB クローリング・サブシステム（以下 WCS:Web Crawling Subsystem）で新規登録ドメインの WEB サイトを自動クローリング・HTML をダウンロードしてそれを自動的に読み込んで機械学習により偽ショッピングサイトかどうかを自動判定することにより，インターネット空間で日々，新規に立ち上がる WEB サイトのクローリングから偽ショッピングサイトかどうかの自動判定まで，一気通貫するシステムである（**図 2**）．

本システムの処理の大まかな流れは，WCS が 2 日前の新規登録ドメインデータをダウンロードし，同データから

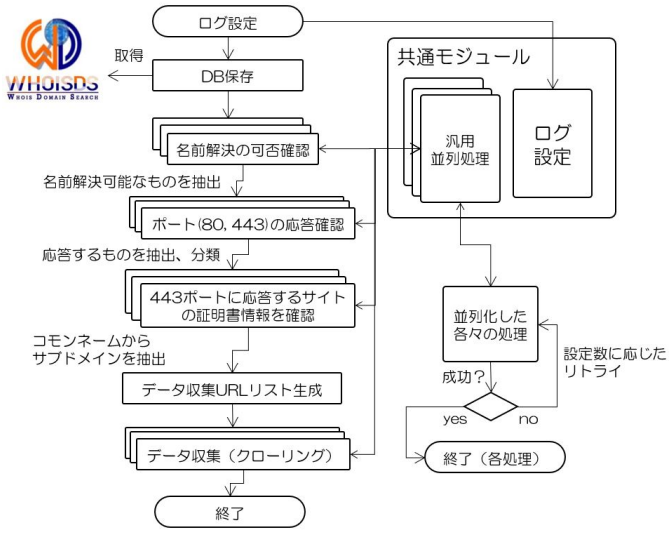

図 3　WCS の設計

Fig. 3　Overview of Web Crawling Subsystem.

生成したリストに基づいて HTML ソースコードを収集し，機械学習サブシステム（以下 MLS:Machine Learning Subsystem）が HTML ソースコードを読み込んで，偽ショッピングサイトの判定を行うものである．

本システムは，CRON バッチにより 1 日に 1 回，深夜の午前 2 時に起動する．詳細は後述するが，まず WEB クローリング・サブシステム（以降 WCS）では WHOISDS から 2 日前の新規登録ドメインをダウンロードし，各ドメインで立ち上がっている WEB サイトの HTML ソースコードをダウンロードする．

ここで収集された新規登録ドメインの WEB サイトの中から新たに設置された偽ショッピングサイトを探し出す訳であるが，1 日にアクセス可能な新規の WEB サイトは数万件という数であり，大半は偽ショッピングサイトには該当しない正規サイト等である．それを人の目で一つ一つ確認して偽ショッピングサイトかどうか判定するにはいくら時間があっても足らない．

そこで，その偽ショッピングサイトかどうかを判定するところを機械学習によって自動化する．こちらも詳細は後述するが，本研究で開発した機械学習・サブシステム（以降 MLS）は，HTML ソースコードを読み込み，ソースコードの特徴から，偽ショッピングサイトかどうかを自動判定するものである．その際，抽出対象は日本消費者をターゲットとした偽ショッピングサイトであるため，日本語のひらがな，カタカナが含まれる WEB サイトのみを対象とするようにした．

3.2　WCS の設計

WCS の設計では，まず先行研究 [33] において半手動で行ったクローリングをプログラムでシーケンシャルに実行するように実装したところ，WHOISDS に登録された新規ドメイン 1 日分のデータをスクレイピングするにあたって，約 1 カ月かかったため，クローリングの処理を名前解決やポートチェック，証明書の情報取集，スクレイピングなどのパーツに分割して，それぞれの処理で適切な並列度を調整可能な並列処理型の Web クローリングシステムを設計した（**図 3**）．

また，データ収集する URL リストは，WHOISDS に登録された情報のままではなく，証明書のコモンネームからサブドメインを抽出するとともに，アクセス先がポート番号 80 と 443 で共通だった場合には 80 番ポートを除外する仕様としている．

なお，システムの実装にあたってはプログラムに汎用性を持たせた共通モジュール化を図っており，同モジュールは後日オープンソースとして公開を予定している．

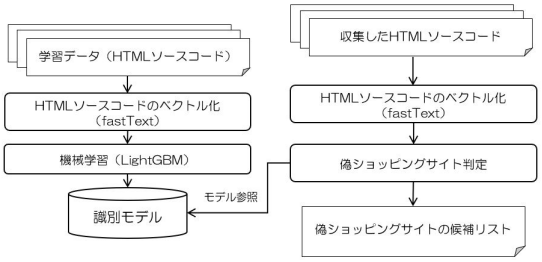

図 4　MLS の概要

Fig. 4　Overview of Machine Learning Subsystem.

3.3　MLS の設計

MLS の設計では，先行研究 [21] で，HTML ソースコードのベクトル化に Doc2Vec，機械学習アルゴリズムに SVM（サポートベクタマシン）を使用していたことに対して，本研究ではベクトル化に fastText，機械学習アルゴリズムに LightGBM(Light Gradient Boosting Machine) を採用した（**図 4**）．WCS で収集された数万件/日という大量のデータを短時間で偽ショッピングサイトか否かを判別するとともに正確性を確保するため，ベクトル化に旧来の Word2Vec に比べて大幅な速度向上が実現されている，米 Facebook 社（現 Meta 社）が開発した自然言語処理ライブラリ fastText を採用し，機械学習アルゴリズムでは機械学習アルゴリズムの自動比較と，パラメータの自動チューニング等を行うことができる AutoML である PyCaret[34] を活用し，Accuracy や処理速度等を比較した結果から LightGBM[35] を採用したものである（**表 1**）．

4.　評価と考察

4.1　WCS の評価

本研究で実施した Web クローリングにおける，WHOISDS で取得したドメイン数とデータを取得できた件数とデータ取得にかかった時間を**表 2** に示す．

本研究の Web Crawling における課題である速度については，データが示されている Zowalla らの研究 [36] を例と

表 1　機械学習アルゴリズムの比較結果
Table 1　Comparison results of machine learning algorithms.

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT(Sec)
Light Gradient Boosting Machine	0.9936	0.9990	0.9981	0.9867	0.9924	0.9868	0.9869	0.3270
Extra Trees Classifier	0.9932	0.9990	0.9999	0.9841	0.9919	0.9860	0.9861	0.2270
K Neighbors Classifier	0.9922	0.9973	0.9946	0.9869	0.9907	0.9840	0.9840	0.5140
Random Forest Classifier	0.9914	0.9989	0.9996	0.9802	0.9898	0.9823	0.9825	0.8200
Gradient Boosting Classifier	0.9877	0.9984	0.9923	0.9787	0.9854	0.9748	0.9749	4.0950
SVM - Linear Kernel	0.9773	0.0000	0.9895	0.9576	0.9733	0.9536	0.9540	0.0240
Decision Tree Classifier	0.9772	0.9757	0.9665	0.9789	0.9726	0.9531	0.9532	0.2190
Linear Discriminant Analysis	0.9765	0.9958	0.9880	0.9571	0.9723	0.9518	0.9522	0.1390
Ada Boost Classifier	0.9753	0.9968	0.9718	0.9693	0.9705	0.9493	0.9493	0.7420
Ridge Classifier	0.9751	0.0000	0.9878	0.9543	0.9708	0.9492	0.9496	0.0190
Logistic Regression	0.9744	0.9954	0.9805	0.9591	0.9697	0.9475	0.9477	0.2610
Quadratic Discriminant Analysis	0.9719	0.9792	1.0000	0.9371	0.9675	0.9428	0.9444	0.1080
Naive Bayes	0.9439	0.9795	1.0000	0.8819	0.9372	0.8869	0.8927	0.0200

表 2　Web クローリング結果
Table 2　Web crawling results

date	targets	result	time
08/03	50,230	34,726	9:34
08/04	29,237	19,943	5:47
08/05	60,639	43,980	11:10
08/06	128,157	87,420	26:12
08/07	99,563	68,025	20:22
08/08	69,836	48,490	13:35
08/09	48,166	32,848	9:34
08/10	54,090	37,347	10:28
08/11	44,035	30,533	7:59
08/12	77,966	53,955	14:56
合計	661,919	457,267	129:37
平均/日	66,192	45,727	12:57

して取り上げて比較すると（**表 3**），Zowalla らが使用した環境が約 32 倍（CPU コア数が 4 倍, メモリが 8 倍）のスペックを持つと仮定して，本研究の速度を 32 倍して比較すると，本研究で開発したクローラは，Zowalla らの Web Crawling システムよりも 3 4 倍程度の速度が出るものと見込まれる．

表 3　Zowalla らの研究と本研究の Web クローリング結果比較
Table 3　Comparison of Web crawling results between other research and this research.

-	Zowalla らの研究	本研究
仮想マシン台数	22	1
CPU	Intel Xeon E5-2689	Intel Xeon E-2324G
CPU コア数	8	2
メモリ	256GB	32GB
データ取得速度	7〜10/sec	0.98/sec
データ取得率	19.76%	69.08%

4.2　MLS の評価

MLS の評価については，正確性と処理スピードの 2 点について実験を行って評価した．

まず，正確性の評価については，実験データとして，実際にインターネット上で公開されていた偽ショッピングサイトのソースコード 1000 件，それと偽ショッピングサイトではない正規サイトのソースコード 1000 件を実際に MLS に自動判定させて，フォールスポジティブ，フォールスネガティブが無いかどうかを測定した．実験を行った結果，フォールスポジティブ 4 件 (正確性 99.6%)，フォールスネガティブ 26 件 (正確性 97.4%) であり，高い判定精度が認められた．

次に，処理スピードの評価は WCS にて取得されたソースコードを実際に機械学習で判定させ，どれくらいの時間で実際に偽ショッピングサイトが抽出されるかを実験した（**表 4**）．

「入力ソース数」が WCS により新規登録ドメインのサイトをクローリングしてダウンロードされた全てのサイトのソースコードであり，当然これらは偽ショッピングサイトでないサイトが大半である．「偽サイト検知数」は，それら大量のソースコードを判定した結果，実際に偽ショッピングサイトとして判定・抽出された数である．以上の実験から，非常に高い精度で，1 日の新規登録ドメインが数万件あろうとその中から高速な処理スピードで偽ショッピングサイトを抽出できる，実用化可能なシステムであることが確認できた．

4.3　偽ショッピングサイト自動検出システム全体の評価

偽ショッピングサイト自動検出システム全体の評価について述べる．これまで WCS と MLS の評価についてそれぞれ述べてきたが，本システムの目的は，あくまでサイバー空間における偽ショッピングサイト脅威の無効化，低

表 4　機械学習処理時間
Table 4　Machine learning processing time

date	入力ソース数	偽サイト検知数	処理時間
08/03	34,726	11	0:36
08/04	19,943	25	0:14
08/05	43,980	5	0:35
08/06	87,420	21	0:59
08/07	68,025	12	0:43
08/08	48,490	9	0:35
08/09	32,848	6	0:33
08/10	37,347	11	0:26
08/11	30,533	94	0:21
08/12	53,955	9	0:37
合計	457,266	203	5:39
平均/日	45727	20.3	0:33

減に寄与することである．そのため，偽ショッピングサイトの探索，セキュリティ事業者等への情報提供を業務とするサイバーパトロール員の業務負担軽減・合理化と工数削減の 2 点につき，本システムの有効性を評価した．

4.3.1　業務負担軽減・合理化に関する評価

既存の業務フローでは，まずサイバーパトロール担当者がインターネット検索エンジンを使用して，偽ショッピングサイトで頻出のキーワード「激安」,「送料無料」等で検索を行い，検索結果から偽ショッピングサイトへ繋がりそうなリンクが見つかるまで，キーワードを変えながら手動で検索を行う．偽ショッピングサイトに繋がりそうなリンクがあれば，リンク先を目視確認し，偽ショッピングサイトによくある特徴である「怪しい日本語」「よく記載されている文言」「価格が異常に安い」に合致するか否かを判定をする．こういった多くの手順を踏みながら，確認・精査の作業を一つ一つ慎重に行い，偽ショッピングサイトと判定したリストを作成し，最終的にはセキュリティ事業者等へ提供していく流れとなる．よって，このように手動で偽ショッピングサイトを探し出していくには，多大な労力と時間がかかるしサイバーパトロールを実施する担当者にも偽ショッピングサイトに関する多くの知見が必要となる．

本研究の成果である新業務フローでは，検索エンジンを使用するのではなく，偽ショッピングサイト検出システムが新規登録ドメイン情報を元に定時処理で自動的に偽ショッピングサイトの候補リストを出力するため，サイバーパトロールの担当者は，リストを確認するだけである．以上から，サイバーパトロール員の作業手順が大きく削減されており，本システムを実用化することで業務負担軽減・合理化に繋がることが期待できる（**図 5**）.

4.3.2　工数削減に関する評価

既存業務フローの作業工数をとある官公庁にヒアリングしたところ，サイバー学生ボランティアから 1 日に約 10 件の偽ショッピングサイトを見つけて報告を受けており，

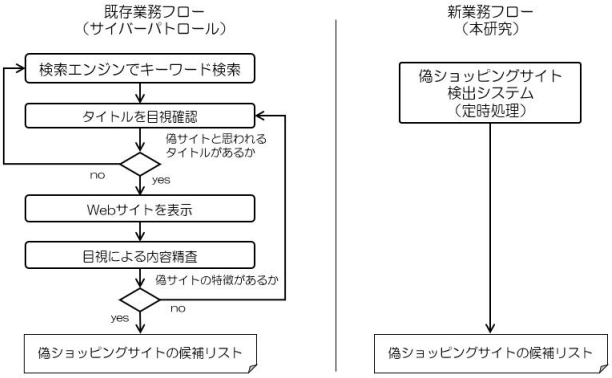

図 5　業務フロー比較
Fig. 5　Task flow comparison.

表 5　工数削減に関する評価
Table 5　Work efficiency evaluation.

手法	1 月あたり工数
既存業務フロー	5.00 人日
新業務フロー	0.83 人日

サイバーボランティアは約 2 時間/日のサイバーパトロール等を行っており，内訳は，サイバーパトロールに約 1 時間 40 分，結果のとりまとめに約 10 分，報告に約 10 分であり，工数にすると，5.00 人日/月となる．これを偽ショッピングサイト自動検出システムを活用した新業務フローで工数を見積もると，サイバーパトロールにかかる時間がまるまる削減されるので，作業時間は結果のとりまとめと報告の 20 分となり，工数は約 0.83 人日/月となることから，本システムを活用することにより，サイバーパトロール員の工数も大幅に削減できるものと評価できる（**表 5**）.

4.3.3　全体評価

以上の評価の結果，本システムを実用化することにより，1 日に 2 時間かけていたサイバーパトロールに要する工数を大幅に削減することができ，サイバーパトロール員の業務負担軽減・合理化に充分寄与しうるシステムであることが確認できた．また，手動で探し出すより，より多くの偽ショッピングサイトを自動検出することに成功したことから，サイバー空間における偽ショッピングサイト脅威の低減・無効化に，より寄与することが期待される．

4.4　考察

4.4.1　WCS の評価に対する考察

Web Crawling の速度について Zowalla らの研究 [36] と比較した WCS の評価では，

- 取得対象とする Web サイト群の応答状況によって速度が大きく変化する
- CPU はコア数以外にもクロック数や用途で性能に違いが出る

- サーバ以外の環境条件が異なる

といった問題があるものの，Web Crawling の研究において速度は一般的な評価指標でないことから，速度についてデータを示されている論文が少ないことから，サブシステムとして概算での評価を行ったものである．

4.4.2 MLS の評価に対する考察

本研究において LightGBM と fastText を活用した機械学習モデルを構築し，実際に偽ショッピングサイトか否かを自動判定させた結果，フォールスポジティブ 4 件，フォールスネガティブ 26 件であり，一部誤判定が見受けられるものの，高い精度を誇ることが確認された．正確性については，誤判定したサイトを学習させる等，更に機械学習モデルをリファインすることにより，改善可能と思われる．また，実際に偽ショッピングサイトか否かを自動判定するための高い精度かつ高速な処理スピードが確認され，実用に耐えうるシステムであると考える．

4.4.3 偽ショッピングサイト自動検出システムの評価に対する考察

偽ショッピングサイト自動検出システムのコンセプトが新規登録ドメインを対象に偽ショッピングサイトを検知することであり, 既存の偽ショッピングサイトに対するアプローチと併用することで新規サイトと既存サイトの両面から偽ショッピングサイトを減らすアプローチを図ることができると推測される．また，既存の手法により収集される偽ショッピングサイトは，検索エンジンによる検索から誘導される偽ショッピングサイトに偏るものであり，本論文における提案手法により，広告から誘導される偽ショッピングサイト等，より幅広いタイプの偽ショッピングサイトを収集できることが期待される．さらに，官公庁でサイバーパトロールに費やしていた工数を偽ショッピングサイト対策の広報・啓発活動や犯罪者の分析，サイトのテイクダウン活動等に使うことで，より高度かつ本質的な活動に集中できる効果も期待される．

5. おわりに

本研究では，偽ショッピングサイト自動検出システムを構築した結果，新規登録ドメインから取得できる HTML ソースコード数が 1 日あたり平均約 4 万 5 千件と飛躍的に向上させることができた．また，大量の HTML ソースコードを機械学習サブシステムにより，高速かつ高い精度で，偽ショッピングサイトかどうか判定させることに成功した．それによって 1 日あたり約 20 件の偽ショッピングサイトを自動抽出することによりインターネット空間の自動クローリングから Web サイトの自動判定まで，一気通貫するシステムを構築したことにより，偽サイトの探索から判定までの業務合理化，工数大幅削減することに成功し，サイバー空間における偽ショッピングサイト脅威の低減・無効化のための活動に十分実用化可能なシステムを構築したものである．

今後は，本手法をフィッシングサイトやテクニカルサポート詐欺サイト等の他の悪質サイトにも適用し，更に幅広く，素早く悪質サイトを検知してサイバー空間における脅威や被害の低減を目指したい．

参考文献

[1] 一般社団法人セーファーインターネット協会, https://www.saferinternet.or.jp (visited on 2022-07)

[2] 一般財団法人日本サイバー犯罪対策センター（JC3：Japan Cybercrime Control Center）, 悪質なショッピングサイト等に関する統計情報（2021年）, https://www.jc3.or.jp/threats/topics/article-431.html (visited on 2022-08)

[3] S.Carolin Jeeva, Elijah Blessing Rajsingh, Intelligent phishing url detection using association rule mining, Human-centric Computing and Information Sciences(2016-07)

[4] Egon Kidmose, Matija Stevanovic, Jens Myrup Pedersen, Detection of Malicious domains through lexical analysis, International Conference on Cyber Security and Protection of Digital Services (Cyber Security).IEEE(2018-06)

[5] MANUEL SÁNCHEZ-PANIAGUA, EDUARDO FIDALGO FERNÁNDEZ, ENRIQUE ALEGRE, WESAM AL-NABKI, AND VÍCTOR GONZÁLEZ-CASTRO, Phishing URL Detection: A Real-Case Scenario Through Login URLs, IEEE Access (Volume: 10)(2022-04)

[6] Mahmoud Khonji, Youssef Iraqi, Andrew Jones, Lexical URL Analysis for Discriminating Phishing and Legitimate Websites, CEAS11:Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference.ACM(2011-09)

[7] Rakesh Verma, Avisha Das, What’s in a URL: Fast Feature Extraction and Malicious URL Detection, IWSPA 17: Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics.ACM(2017-03)

[8] Vara Vundavalli, Farhat Barsha, Mohammad Masum, Hossain Shahriar, Hisham Haddad, Malicious URL Detection Using Supervised Machine Learning Techniques, SIN 2020:13th International Conference on Security of Information and Networks.ACM(2020-11)

[9] Shantanu, B Janet, R Joshua Arul Kumar, Malicious URL Detection: A Comparative Study,2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS).IEEE(2021-03)

[10] Aaron Blum, Brad Wardman, Thamar Solorio, Gary Warner, Lexical feature based phishing URL detection using online learning, Proceedings of the 3rd ACM workshop on Artificial intelligence and security.ACM(2010-08)

[11] Ram B. Basnet, Andrew H. Sung, Quingzhong Liu, Feature Selection for Improved Phishing Detection, AProceedings of the 25th international conference on Industrial Engineering and Other Applications of Applied Intelligent Systems: advanced research in applied artificial intelligence(2012-06)

[12] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri, Machine learning based phishing detection from URLs, Expert Systems With Applications(2018-09)

[13] Thuy Thi Thanh Pham, Van Nam Hoang, Thanh Ngoc Ha, Exploring Efficiency of Character-level Convolution Neuron Network and Long Short Term Memory on Ma-

licious URL Detection, ICNCC 2018:Proceedings of the 2018 VII International Conference on Network, Communication and Computing.ACM(2018-12)

[14] Nabeel Al-Milli, Bassam H. Hammo, 11th International Conference on Information and Communication Systems (ICICS).IEEE(2020-04)

[15] Buket Geyik, Kubra Erensoy, Emre Kocyigit, Detection of Phishing Websites from URLs by using Classification Techniques on WEKA,Sixth International Conference on Inventive Computation Technologies (ICICT) IEEE(2021-01)

[16] Mohammed Abutaha, Mohammad Ababneh, Khaled Mahmoud, Sherenaz Al-Haj Baddar, URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis, 12th International Conference on Information and Communication Systems(ICICS).IEEE(2021-05)

[17] Guang Xiang, Jason Hong, Carolyn P. Rose, Lorrie Cranor, CANTINA+: A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites, ACM Transactions on Information and System SecurityVolume 14Issue 2 Article No.: 21pp 1–28(2011-09)

[18] Jia-Chang Xu, Kilho Shin, Yu-Lu Liu, Detecting Fake Sites based on HTML Structure Analysis, ICCNS 16: Proceedings of the 6th International Conference on Communication and Network Security Pages:86–90(2016-11)

[19] Taichi Ishikawa, Yu-Lu Liu, David Lawrence Shepard, Kilho Shin, Machine Learning for Tree Structures in Fake Site Detection, ARES 20: Proceedings of the 15th International Conference on Availability, Reliability and Security.ACM(2020-08)

[20] Jian Feng, Yuqiang Qiao, Ou Ye, Ying Zhang, Detecting phishing webpages via homology analysis of webpage structure, PeerJ. Computer Science(2022-02)

[21] Naoki Kurihara, Hidenori Tsuji, Masaki Hashimoto, Spoofed Website Detection using Machine Learning, IEICE technical report vol.118, no.315, ICSS2018-56, pp.19-24(2018-11)

[22] Ankit Kumar Jain, B. B. Gupta, A machine learning based approach for phishing detection using hyperlinks information, Journal of Ambient Intelligence and Humanized Computing volume 10, pages2015–2028 (2019-10)

[23] Syed Rameem Zahra, Mohammad Ahsan Chishti, Asif Iqbal Baba, Fan Wu, Detecting Covid-19 chaos driven phishingmalicious URL attacks by a fuzzy logic and data mining based intelligence system, Egyptian Informatics Journal Volume 23, Issue 2, Pages 197-214(2022-07)

[24] Mohsen Sharifi, Seyed Hossein Siadati,A phishing sites blacklist generator, IEEE/ACS International Conference on Computer Systems and Applications(2008-04), IEEE(2008-04)

[25] Mohammed Nazim Feroz, Susan Mengel, Phishing URL Detection Using URL Ranking, IEEE International Congress on Big Data(2015-08)

[26] Rami M.Mohammad, Fadi Thabtah, Lee McCluske, Intelligent rulebased phishing websites classification, Volume8, Issue3,Pages 153-160, IET Information Security(2014-03)

[27] Yury Zhauniarovich, Issa Khalil, Ting Yu, Marc Dacier, A Survey on Malicious Domains Detection through DNS Data Analysis, ACM Computing Surveys(2018-02)

[28] Ehsan Nowroozi, Abhishek, Mohammad Reza Mohammadi, Mauro Conti, An Adversarial Attack Analysis on Malicious Advertisement URL Detection Framework, arXiv:2204.13172v1 [cs.LG] 27(2022-04)

[29] Christopher Olston and Marc Najork. 2010. Web Crawling. Found. Trends Inf. Retr. 4, 3 (March 2010), 175–246. https://doi.org/10.1561/1500000017

[30] Carlos Castillo. 2005. Effective web crawling. SIGIR Forum 39, 1 (June 2005), 55–56. https://doi.org/10.1145/1067268.1067287

[31] Tchakounte, Franklin Ngnintedem, Jim Irépran, Damakoa Faissal, Ahmadou Fotso, Franck. (2021). Crawl-shing: A Focused Crawler for Fetching Phishing Contents based on Graph Isomorphism. Journal of King Saud University - Computer and Information Sciences. 10.1016/j.jksuci.2021.11.003.

[32] Sotiris Batsakis, Euripides G.M. Petrakis, Evangelos Milios, Improving the Performance of Focused Web Crawlers, Data & Knowledge Engineering Volume 68, Issue 10, Pages 1001-1013(2009-10)"

[33] Kazuki Kato, Masaki Hashimoto, 偽サイト検出のための Web クローリングシステムの開発, Master's Thesis, Institute of Information Security(2021-03)

[34] Welcome to PyCaret - PyCaret Official, https://pycaret.org/

[35] Welcome to LightGBM's documentation!, https://lightgbm.readthedocs.io/

[36] Richard Zowalla, Thomas Wetter, Daniel Pfeifer, Crawling the German Health Web: Exploratory Study and Graph Analysis, Journal of Medical Internet Research, 22(7):e17853(2020-07)