HTMLタグの構造に着目したグラフ畳み込みネットワークによる悪性サイト判定

山本 貴巳[1,a)]　Kim Sangwook[1]　班 涛[2]　高橋 健志[2]　小澤 誠一[1]

概要：近年，Drive-by Download 攻撃やフィッシングといった Web 媒介型攻撃による被害が深刻になっており、直接サイバー攻撃を仕掛けたり，攻撃につながるウェブサイトに誘導したりする悪性サイトの迅速な発見が求められている．これまで，URL に着目して，文字レベル畳み込みニューラルネットの学習や URL に頻出する単語を許可リスト化，無視リスト化することで，悪性サイトの判定システムを開発してきた．本研究では，この判定システムの拡張として，HTML タグの構造に着目したグラフ畳み込みネットワークによる悪性サイト判定の手法を提案する．HTML 文書は Document Object Model (DOM) によって表現することができ，HTML タグの構造を木構造として表現できる．これを HTML 文書の「グラフ表現」と呼び，グラフ畳み込みネットワークを適用して悪性サイト判定を行う方法を提案する．従来の URL を利用した悪性サイト判定システムに，HTML タグの構造による判定を加えることで，高い精度で悪性サイトの検知が可能であることを示す．

キーワード：サイバーセキュリティ，機械学習，ウェブ媒介型攻撃，悪性サイト判定，HTML コンテンツ

Malicious-Site Detection Using Graph Convolutional Networks Based on HTML Tag Structure

Takami Yamamoto[1,a)]　Sangwook Kim[1]　Tao Ban[2]　Takeshi Takahashi[2]　Seiichi Ozawa[1]

Abstract: Recently, damage caused by web-based attacks such as drive by-download attacks and phishing has become serious. There is a pressing need to quickly detect malicious websites which directly launch cyberattacks or guide users to dangerous sites to trigger attacks. In previous work, we have developed a malicious-site detection system based on the word frequency of URLs using character-level convolutional neural networks. In this research, as an extension of the existing system, we propose a new approach using graph convolutional networks based on structural features of HTML tags. First, an HTML document is represented by the document object model (DOM) as a graph to captures structural information in the document. Then, malicious-site detection is performed by applying graph convolutional networks to the graphs. We show that malicious sites can be detected with higher accuracy by incorporating structural information carried by the HTML tags than the previous approach based on URL.

Keywords: Cyber security, Machine learning, Web-based attacks, Malicious-site detection, HTML content

1. はじめに

近年，Drive-by Download 攻撃やフィッシングといった Web 媒介型攻撃による被害が深刻になっており、直接サイバー攻撃を仕掛けたり，攻撃につながるウェブサイトに誘導したりする悪性サイトの迅速な発見が求められている．

[1] 神戸大学
Kobe University
[2] 情報通信研究機構
National Institute of Information and Communications Technology
[a)] 202t263t@stu.kobe-u.ac.jp

従来から利用されている悪性サイトへの対策方法として，Google Safe Browsing（GSB）[1] などの無視リストが用いられている．しかし，無視リストでは既知の悪性サイトしか検知できず，未知の悪性サイトに対して脆弱であるという問題がある．そこで，未知の悪性サイトも検知するために，機械学習を利用した悪性サイト検知に関する研究が盛んである．機械学習を利用するのは，悪性サイトに潜む特徴を学習して汎化性能を得ることで，未知の悪性サイトに対しても検知が可能となるからである．

これまで，NICT 委託研究である WarpDrive プロジェクト [2] における研究の一環として，URL 文字列から得られる情報を用いた悪性サイト判定システムを開発してきた．具体的には，URL 文字列から悪性サイトかどうか判定する Character-Level CNN（CL-CNN）モデル [3] と，URL に頻出する単語群から作成した許可リスト，無視リストを組み合わせることで悪性サイト検知を行う [4]．この従来手法では URL 文字列から得られる情報のみで悪性サイトの判定を行っているが，Web サイトから得られる情報として他にも HTML 文書がある．本研究では，従来手法の拡張として，HTML タグの構造に着目したグラフ畳み込みネットワークによる悪性サイト判定の手法を提案する．HTML 文書は Document Object Model（DOM）[5] によって表現することができ，HTML タグの構造を木構造として表現できる．これを HTML 文書の「グラフ表現」と呼び，グラフ畳み込みネットワークを適用して悪性サイト判定を行う方法を提案する．WarpDrive で収集したデータを使用して実験を行ったところ，641 件ある悪性サイトの約 7 割に相当する 447 件を正しく検知することができた．

2. 先行研究

本論文で提案する手法の前に，従来手法である URL を利用した悪性サイト判定システム [4] について説明する．この悪性サイト判定システムでは，Web サイトの危険度をグレーリストとダークグレーリストの二段階にわけて判定している．本節では，グレーリストとダークグレーリストの判定方法について説明したのちに，従来手法による悪性サイト判定の結果について説明する．

2.1 グレーリスト

グレーリストはある程度の危険度があると判定された Web サイト群のことで，許可リストと CL-CNN モデルにより判定される．この許可リストは良性サイトの URL に頻出する単語群から成るもので，許可リストの単語を含む URL を良性 URL として扱う．SentencePiece[6] を用いて URL の分かち書きを行い，各単語の出現頻度を計算して閾値より高いものを許可リストの単語として抽出する．CL-CNN は画像認識で使用される CNN を自然言語処理に適応させたもので，その汎用性の高さから URL 文字列による悪性サイト判定にも利用できることがわかっている [7]．悪性サイト判定の場合，URL を一文字ずつ区切り Embedding 層でベクトルへと変換してから，通常の CNN と同じように畳み込み層とプーリング層により特徴抽出を行う．グレーリストの作成は次の手順で行う．初めに，分析の対象となる Web サイト群に対して許可リストを適用して良性サイトを取り除く．そして，残りの Web サイト群を CL-CNN モデルで判定して，悪性と判定されたものをグレーリストとする．

2.2 ダークグレーリスト

ダークグレーリストはグレーリストの中でもより危険度が高いと判定された Web サイト群のことである．悪性サイトを作る攻撃者は URL を作成する時に，ユーザのアクセスを誘導するような仕掛けを用意するケースがある．例えば，‘download’ という単語はダウンロードサイトであることを示しており，ユーザのアクセスを誘導する意図が有ると考えられる．このように攻撃者が意図をもって作成した URL を検知するために無視リストを導入する．無視リストは悪性サイトの URL に頻出する単語群から成り，2.1 節の許可リストと同様に各単語の出現頻度を計算して，閾値より高いものを無視リストの単語として抽出する．

他にも，ユーザのアクセスを誘導するような URL として，フィッシングサイトによく見られるホモグラフ攻撃がある．ホモグラフ攻撃とは正規サイトのドメインを偽装して罠サイトに誘導する攻撃のことである．例えば，‘amazno’ や ‘amazzon’ などの文字列を罠サイトのドメインに含ませておくことで，ユーザに ‘amazon’ の Web サイトと勘違いさせてアクセスを誘導することができる．ホモグラフ攻撃を検知するためには，攻撃の対象となりやすい正規サイトのドメインとの類似度を測る必要がある．そこで，fastText[9] を用いて文字列を Embedding 空間で表現することで，正規サイトのドメインとの類似度を計算する．以上がダークグレーリストの判定方法であり，無視リストの単語か，正規サイトのドメインとの類似度が高い文字列を含むグレーリスト内の Web サイトを，より危険度が高いダークグレーリストとする．

2.3 従来手法による悪性サイト判定の結果

WarpDrive ではユーザ参加型のサイバー攻撃観測網を構築しており，専用のアプリケーションを通じてユーザがアクセスした Web サイトの URL を収集している．2019 年 2 月から 2020 年 7 月までに収集された URL から，FQDN の重複を除いたものをテストデータとした．そして，従来手法でグレーリストとダークグレーリストの作成を行った結果が**表 1** になる．また，ダークグレーリストの URL について，その FQDN を Web サイトのマルウェア検査サービスである Virus Total[10] で検索したときの，良性サイト

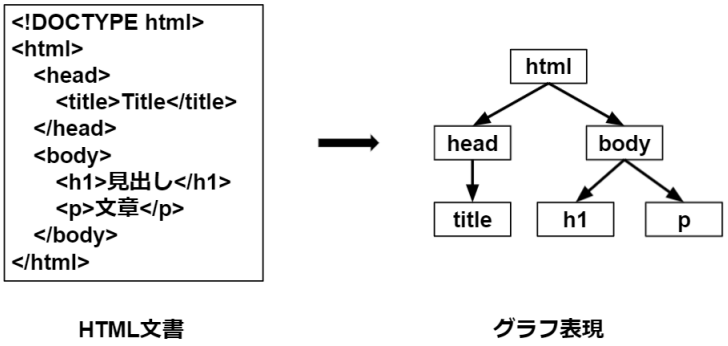

図 1 HTML 文書のグラフ表現
Fig. 1 Graph of HTML documents.

表 1 グレーリストとダークグレーリスト
Table 1 Graylist and Darkgraylist.

	Web サイトの数
テストデータ	2,877,719 (100%)
グレーリスト	265,239 (9.2%)
ダークグレーリスト	134,218 (4.6%)

表 2 ダークグレーリストの内訳
Table 2 Breakdown of Darkgraylist.

	Web サイトの数
良性サイト	128,954 (96.1%)
悪性サイト	5,264 (3.9%)

と悪性サイトの内訳が**表 2** になる．約 290 万件近くあるテストデータから危険度が高い Web サイトを絞り込めてはいるものの，まだ偽陽性が多いという問題がある．この結果から，URL 文字列から得られる情報だけでは悪性サイト判定の精度に限界があると言える．したがって，より確度の高い悪性サイト判定を実現するには，URL 以外の Web サイトに関する情報を利用する必要がある．本研究では，従来手法の URL による判定から得られたダークグレーリストをもとに，HTML 文書による判定を加えることで悪性サイト判定システムの性能がどれほど向上するかを検証する．

3. 提案手法

本研究では，従来の悪性サイト判定システムの拡張として，HTML タグの構造に着目したグラフ畳み込みネットワークによる悪性サイト判定の手法を提案する．まず，3.1 節では HTML 文書をグラフデータに変換する方法について説明し，3.2 節では Graph Convolutional Networks（GCN）の利用方法について説明する．

3.1 HTML 文書のグラフ表現

HTML 文書を GCN の入力データにするにはグラフデータに変換する必要がある．マークアップ言語である HTML は，HTML タグを使用することで HTML 文書の構造を指定したり，文字に様々な効果を持たせることができる．また，HTML 文書を簡単に操作可能にするためのデータ表現の一つとして Document Object Model（DOM）というものがあり，HTML タグの階層構造にしたがって，HTML 文書を木構造として表現することができる．HTML 文書のグラフ表現の例を**図 1** に示す．HTML 文書のグラフ表現では，HTML タグの名前をノードに割り当て，タグで囲まれたテキストは扱わないものとする．ノードに割り当てる HTML タグには，HTML5 に用意されている 107 個の HTML タグ [11] を使用し，それ以外の HTML タグに対しては ‘unknown’ を割り当てる．[12] によれば，Web サイトの 88.7%で HTML5 が使用されており，ほとんどの Web サイトに対応することができる．また，HTML タグの階層構造がノード間の親と子の関係に対応している．このように HTML 文書をグラフ表現に変換することで，GCN による学習が可能となる．

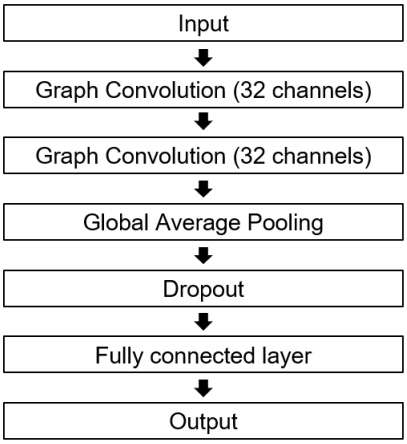

図 2 GCN モデル
Fig. 2 GCN model.

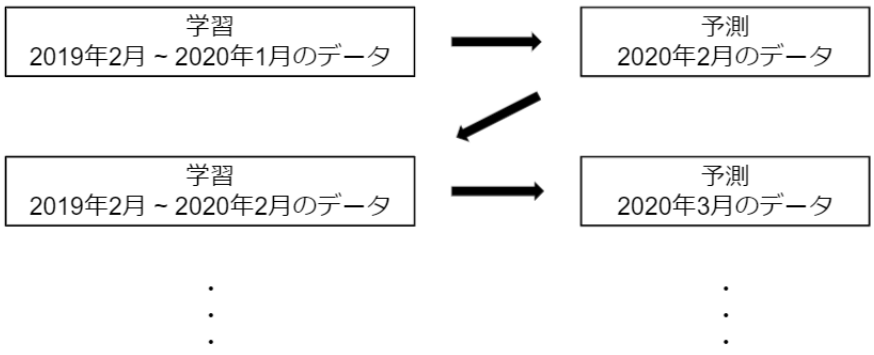

図 3 実験の流れ
Fig. 3 Experiment cycle.

3.2 Graph Convolutional Networks

GCN とは，グラフニューラルネットワークに CNN で利用される畳み込み演算を適用したものである．グラフデータにおける畳み込み演算は，あるノードが持つ特徴量に対して，そのノードと隣接しているノードの特徴量に重みをかけたものを足し合わせるという処理である．この畳み込み演算をグラフ上で行うことで，あるノードにはどのようなノードが隣接しているかという情報も考慮することができる．畳み込み演算の方法については様々な手法が提案されているが，本研究では Kipf ら [13] による手法を利用する．通常の GCN では，隣接しているノードも考慮した各ノードの特徴量は得られるが，グラフ分類を行うにはグラフ自体の特徴量も必要である．そこで，プーリング層で全ノードの特徴量の平均値を取り，その値をグラフの特徴量とする．

3.1 節で得られた HTML 文書のグラフ表現をそのまま GCN の入力にすることはできないので，グラフ表現を数値に変換する必要がある．グラフ表現のノードには HTML タグの名前が割り当てられており，これを One-Hot ベクトルで表すことでノードの特徴量とする．HTML タグには ‘unknown’ も含めて 108 個を想定しているので，ノードの特徴量は 108 次元となる．エッジの有無は隣接行列で表し，図 1 のように有向グラフとする．また，エッジ自身に特徴量はないものとする．

最後に，提案手法で用いた GCN モデルの構成を**図 2** に示す．畳み込み層は二層で，活性化関数には ReLU を使用した．

4. 評価実験

4.1 データセット

本研究では，2 章の従来手法から得られたダークグレーリストをデータセットとして提案手法の性能評価を行った．ダークグレーリストの URL から，Python の HTTP ライブラリである Requests[14] を使用して該当 Web サイトの HTML 文書を収集した．2021 年 8 月時点で収集できた HTML 文書の数とその良性サイト，悪性サイトの内訳を**表 3** に示す．Web サイトのラベル付けは VirusTotal での検索結果を利用した．

表 3 データセットの内訳
Table 3 Breakdown of dataset.

	HTML 文書の数
良性サイト	51,578 (97%)
悪性サイト	1,586 (3%)
合計	53,164 (100%)

表 4 予測精度
Table 4 Prediction accuracy.

正解率	適合率	再現率	F 値
0.564	0.041	0.697	0.077

表 5 混同行列
Table 5 Confusion Matrix.

		予測値	
		悪性	良性
真値	悪性	447	194
	良性	10431	13274

4.2 提案手法による悪性サイト判定の結果

実験で使用する HTML 文書のデータセットは 2019 年 2 月から 2020 年 7 月に WarpDrive で収集された URL をもとに作成しており，時系列データになっている．そこで，データセットを一月分ごとにわけて**図 3** のように学習と予測を繰り返して実験を行った．図 3 では，各月のテストデータに対して，それ以前のデータを全て学習データとして GCN モデルを学習して予測をしている．テストデータには 2020 年の 2 月から 7 月までのデータを使用するので，学習と予測を 6 回繰り返すことになる．

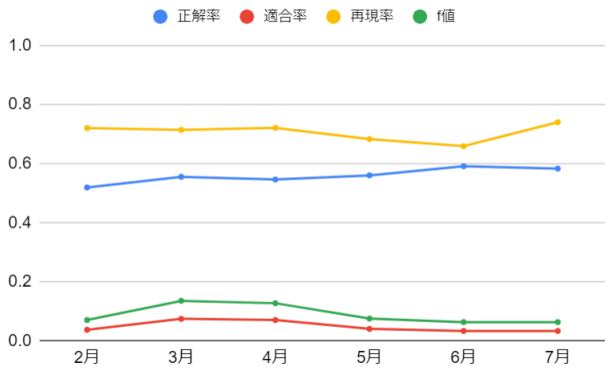

図 4 月ごとの実験結果
Fig. 4 Monthly experimental results.

実験結果について，全テストデータに対する予測精度と混同行列を**表 4**，**表 5** に示し，2020 年 2 月から 7 月までの各月のテストデータに対する予測精度の推移を**図 4** に示す．表 4 から，再現率は約 0.7 で 641 件の悪性サイトの内，447 件を正しく検知できていることがわかる．また，図 4 では，テストデータの全期間に渡って再現率が常に 0.7 近くあることを示しているが，逆に言うとどの月も 2 割から 3 割の悪性サイトを見逃していることになる．したがって，それまでの悪性サイトと比べて HTML タグの構造が大きく異なる悪性サイトが，一カ月ごとに 2 割から 3 割は発生していると考えられる．また，適合率に大幅な改善は見られなかった．適合率が極端に低い原因の一つとして，実験で使用しているデータが不均衡データである点が挙げられる．現実的な問題を考えた場合，今回使用したデータよりも悪性データに比べて良性データの方が多いという状況もあり得るので，適合率の低さは対処しなければならない問題である．不均衡データの学習については多くの手法が提案されているので，今回のデータと提案手法に適したものを検討する必要がある．

5. 結論

本研究では，従来手法である URL 文字列から得られる情報を用いた悪性サイト判定システムの拡張として，HTML タグの構造に着目したグラフ畳み込みネットワークによる悪性サイト判定の手法を検討した．提案手法では，HTML 文書の特徴量として，HTML タグの構造から得られるグラフ表現を定義し，グラフデータを学習することができる機械学習モデルの GCN を用いて Web サイトの悪性判定を行った．2019 年 2 月から 2020 年 7 月に WarpDrive で収集されたデータを用いて実験を行った結果，641 件ある悪性サイトの約 7 割に相当する 447 件を正しく検知することができた．再現率は常に 0.7 近くある一方で，適合率に大幅な改善は見られなかった．

今後の取り組みでは，HTML タグの名前だけでなく，その中身である HTML 要素や HTML タグに付加された情報である HTML 属性も考慮したグラフ表現の構築方法について検討し，提案手法の改善を図る．HTML 文書自体を一つのグラフとして表現することで，包括的な HTML 文書の悪性判定が可能になる．また，本研究では HTML 文書を扱ったが，Web サイトから得られる情報として他にも JavaScript がある．JavaScript は Web 媒介型攻撃の実行を担う要素であり，悪性サイト判定において重要だと考えられる．JavaScript による悪性サイト判定の手法を従来手法や提案手法と組み合わせることで，悪性サイト判定システムの精度向上を目指す．

謝辞 本研究成果は，科研費基盤 (B)「機械学習とドメイン知識を導入した攻撃生成過程のモデル化と実データによる検証」（16H02874）により得られたものです．本研究を進めるにあたり，数々の有益なご教授を賜りました研究メンバーの皆様に心より感謝申し上げます．

参考文献

[1] `https://safebrowsing.google.com/`, 2021 年 8 月 22 日閲覧.
[2] `https://warpdrive-project.jp/`, 2021 年 8 月 22 日閲覧.
[3] Xiang Zhang, Junbo Zhao, and Yann LeCun, Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pp. 649–657. Curran Associates, Inc. (2015)
[4] 山本 貴巳, 石川 真太郎, 山田 明, 小澤 誠一, 文字レベル畳み込みニューラルネットによる悪性サイト判定の URL 単語頻度に基づく高度化. 暗号と情報セキュリティシンポジウム (2021)
[5] `http://www.w3.org/TR/2004/REC-DOM-Level-3-Core-20040407/core.html`, 2021 年 8 月 22 日閲覧.
[6] Taku Kudo and John Richardson, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv:1808.06226* (2018)
[7] Joshua Saxe and Konstantin Berlin, eXpose: A character-level convolutional neural network with embeddings for detecting malicious URLs, file paths and registry keys. *arXiv:1702.08568* (2017)
[8] Hung Le, Quang Pham, Doyen Sahoo, Steven C.H. Hoi, URLNet: Learning a URL representation with deep learning for malicious URL detection. *arXiv:1808.03162* (2018)
[9] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*. 5, 135-146 (2017)
[10] `https://www.virustotal.com/gui/home/url`, 2021 年 8 月 22 日閲覧.
[11] `http://www.htmq.com/`, 2021 年 8 月 22 日閲覧.
[12] `https://w3techs.com/technologies/details/ml-html5`, 2021 年 8 月 22 日閲覧.
[13] Thomas N. Kipf, Max Welling, Semi-Supervised Classification with Graph Convolutional Networks. *arXiv:1609.02907* (2016).
[14] `https://requests-docs-ja.readthedocs.io/en/`

latest/, 2021 年 8 月 22 日閲覧.