6V-07

DNS 情報を用いた不正 Web サイト検知システムの提案

金澤 しほり† 中村 嘉隆† 高橋 修†

†公立はこだて未来大学 システム情報科学部

1. はじめに

近年，Web サイトを利用した攻撃が急増している．ユーザを偽のサイトへ誘導するフィッシングや、悪意のあるプログラムをダウンロードさせるガンブラー攻撃法などが代表的なものであるが，これらの攻撃法により，個人情報等を不正に取得され，経済的な被害を受ける事件が増加している．警察庁広報資料のインターネットバンキングに係る不正送金事犯の発生状況によると，2013 年 5 月から 11 月までの間に発生件数が 250 件近くまで急増しており，現在も増加傾向にある[1]．このような被害を防ぐには，不正 Web サイトを早急に発見し，不正 Web サイトへ誘導する通信を遮断するなどの対策が必要である．これまでに，ファイアウォールや， IDS(侵入検知システム)といった技術が開発され，対策手法として用いられている．しかし，これらの手法は，既知の不正 Web サイトや通信に対して有効であり，未知の不正 Web サイトや通信には対応できないため，これらに対応した仕組みが必要である．

2. 関連研究

不正 Web サイトの検出手法として，DNS ログを参照して，マルウェアに感染されているクライアントからアクセスされたドメインを抽出するもの[2]や，10 文字以上で構成されたドメイン名や，英数字が混在したドメイン名から不正 Web サイトを抽出するもの[3]などが提案されている．しかし，前者は広告サイトなども含まれてしまい，正規 Web サイトと不正 Web サイトの検出精度が低い．後者は未知の不正 Web サイトの検出数が少なく，検出範囲も狭い．このように，従来手法は，誤検出が多く，検出できる不正 Web サイトも限定されているため，実際に運用するためには問題が多い．

3. アプローチ

関連研究の問題点は，検出条件に Domain Name System(DNS)から得られる情報を加えることで解決できる可能性がある． DNS は Web サイトと関連の強い情報を管理しており，不正 Web サイトの正確な検知を行うために最適な情報を保持しているため，DNS 情報内で有効な検出条件を設定することで，検出精度の向上を図ることが可能である．

4. DNS 情報を用いた不正 Web サイトの検知システム

本研究では，DNS 側で得られる情報のうち，ドメイン名と IP アドレスを用いて未知の不正 Web サイトの検出率・検出精度向上を図る．

4. 1 未知の不正 Web サイト抽出方法

未知の不正 Web サイト抽出方法を図 1 に記す．

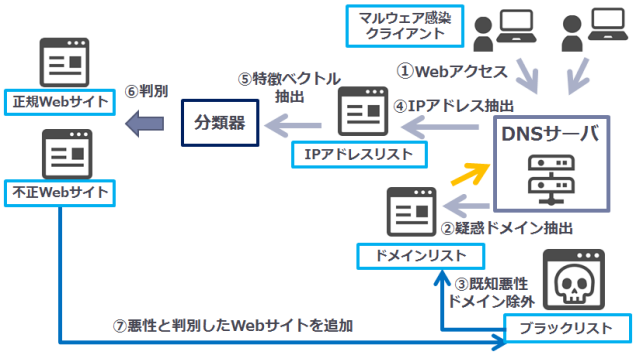

図 1 未知の不正 Web サイト抽出方法

DNS サーバに Web アクセスが行われているクライアントの中から，既知の悪性ドメインにアクセスしているクライアントを，マルウェアに感染されているクライアント(以下マルウェア感染クライアント)として検出する．一般にマルウェアは，感染を拡大させるために多数の不正 Web サイトへアクセスを試みる(図 1 ①)．そのため，不正 Web サイトは，同時に複数のマルウェア感染クライアントからアクセスが行われている可能性が高い．そこで，マルウェア感染クライアントから重複してアクセスが行われているドメインを DNS サーバの中から抽出し(図 1 ②)，ブラックリストと照合して，既知の悪性ドメインを除外し，照合せず残ったドメインを未知の不正 Web サイトの可能性が高いとみて抽出する(図 1 ③)．抽出したドメインが正規 Web サイトか不正 Web サイトであるか判別するため，DNS サーバから IP アドレスを取得して利用する(図 1 ④)．また，IP アドレスを構成するビット列の構造的な性質から，アドレスに固有な特徴ベクトルを抽出し(図 1 ⑤)，教師あり機械学習法の一つである Support Vector Machine(SVM)を用いて該当 IP アドレスの良性・悪性を判別し，不正 Web サイトを検知する(図 1 ⑥)．このとき，IP アドレスの良性・悪性は，良性を正規 Web サイト，悪性を不正 Web サイトとし，特徴ベクトルは，IP アドレスの上位 1～4 オクテットごとにバイナリビット列に変換して構成する．

SVM が不正 Web サイトであると判別した場合は，さらに，抽出した不正 Web サイトが属するネットワークアドレスに探索範囲を拡張して未知の不正 Web サイトを特定する．悪質な活動に利用される IP アドレスにはネットワークブロックの空間的に偏りが現れ，悪性 IP アドレスが密集する傾向があると知られているため[4]，ネットワークアドレスに探索範囲を拡張することで，未知の不正 Web サイトの検出率の向上を図る．

5. 基礎評価実験

IP アドレスを用いて良性・悪性を判別するにあたり，IP アドレスを用いた最大限の精度を引き出すことがで

“A detecting system of malicious web site using information stored in DNS”

Shihori Kanazawa†, Yoshitaka Nakamura†, Osamu Takahashi†

†School of Systems Information Science, Future University Hakodate

きる訓練データの状態を把握するため，SVM を用いた基礎評価実験を行った．

5.1 データ

SVM の訓練データとして，悪性 IP アドレスのリストと，良性 IP アドレスのリストを用いる．悪性 IP アドレスのリストは，Drive-by Download Data by Marionette(D3M)で利用されている IP アドレスを抽出し，クライアント側の IP アドレスと DNS に問い合わせが行われた IP アドレスを除外して利用する．通常の IP アドレスのリストは，Alexa Top Global Sites[5]を用いて抽出した IP アドレスを利用する．

5.2 実験内容

良性・悪性のIPアドレスを5:5の割合で含んだリストを作成し，訓練データ数1000個と2000個に対して5分割交差検定を行った．5分割交差検定では，データを5つに分割し，その中の1つをテストデータ，残りを訓練データとして用いることで評価する．データを先頭から分割していき，順番にテストデータとして評価するため，テストデータにそれぞれ分割番号を割り振った．分割番号1，2のテストデータは全て良性，分割番号3のテストデータは良性・悪性両方が含まれており，分割番号4，5のテストデータは全て悪性である．これらの訓練データを用いた識別精度の平均値と，各オクテットの精度を出力することで，データ数とオクテットの関係が精度に影響を及すのか実験を行った．

5.3 実験結果

データ数 1000 個で行った交差検定の精度を表 1，データ数 2000 個で行った交差検定の精度を表 2，各データ数における精度を表 3 に示す．データ数 1000 個の場合よりも，2000 個の方が，全体的に精度が向上している．データ数 1000 個で行った 5 分割交差検定の各精度を表 2 に，2000 個で行った 5 分割交差検定の各精度を表 3 に示す．データ数 1000 個の分割番号が 1，データ数 2000 個の場合の分割番号が 1，2 の精度は低く，残りのテストデータでは精度が向上した．このとき，データ数 1000 個の場合の分割番号 1，データ数 2000 個の場合の分割番号 1，2 のテストデータは良性であり，残りの分割番号で用いたテストデータは悪性であった．また，データ数 1000 個，2000 個ともに，第 4 オクテットで精度は低下した．

表 1 交差検定の精度(データ数 1000 個)

分割番号	1	2	3	4	5
第 1 オクテット	51.0%	70.0%	59.0%	56.0%	74.5%
第 2 オクテット	44.5%	69.5%	65.0%	58.5%	72.0%
第 3 オクテット	44.5%	65.5%	62.5%	62.0%	66.5%
第 4 オクテット	49.0%	64.5%	63.5%	64.5%	65.5%

表 2 交差検定の精度(データ数 2000 個)

分割番号	1	2	3	4	5
第 1 オクテット	60.8%	49.5%	70.8%	70.8%	68.3%
第 2 オクテット	64.8%	53.5%	75.8%	69.3%	67.0%
第 3 オクテット	59.5%	56.5%	73.3%	67.5%	65.3%
第 4 オクテット	59.8%	60.3%	70.3%	65.8%	62.8%

表 3　識別精度の平均値

オクテット	1	2	3	4
データ数 1000 個	62.1%	61.9%	60.4%	61.4%
データ数 2000 個	64.0%	66.0%	64.3%	63.8%

6. 考察

データ数 2000 個の精度が 1000 個の場合よりも全体的に向上したことから， IP アドレスを訓練データとして用いる際は，より多くの訓練データを利用した方が正確な判別ができると考えられる．また，テストデータが良性の場合よりも，悪性の方に精度の向上が見られることから，提案手法では IP アドレスで判別する前に，ドメイン情報を用いて既知の良性・悪性ドメインを除いた未知のドメインを導き出すが，未知のドメインは悪性の可能性が高いことが判明しているので[2]，悪性の IP アドレスを判別するには有効である．

7. おわりに

本稿では，未知の不正 Web サイトの検出率・検出精度を向上するために，DNS 側から得られる情報のうち，IP アドレスを利用した検出手法と識別方法を提案した．IP アドレスを用いた判別は，判別する対象が悪性の可能性が高いものほど精度が向上することがわかった．

今後は， IP アドレスを用いた未知の不正 Web サイトの検出率を調査し，ドメイン情報を用いた場合の検出率と比較・検討する予定である．

参考文献

[1] 警察庁広報資料，"平成 25 年中のインターネットバンキングに係る不正送金事犯の発生状況"，入手先 https://www.scutum.jp/information/falsification2.html,2013.

[2] 田中晃太郎，長尾篤，森井昌克，"DNS ログからの不正 Web サイト抽出について-解析手法とその匿名-，"コンピュータセキュリティシンポジウム 2013 論文集，Vol.2013, No.4, pp.132-138, 2013.

[3] 劉亦晨，後藤滋樹，"DNS 情報による悪意のあるサイトの検出法，" 2012 年度 早稲田大学大学院 基幹理工学研究科 情報理工学専攻 修士論文

[4] 千葉大紀，森達哉，後藤滋樹，”悪性 Web サイト探索のための優先巡回順序の選定法” コンピュータセキュリティシンポジウム 2012 論文集, Vol.2012, No.3, pp.805-812, 2012.

[5] Alexa Top Sites，http://www.alexa.com/topsites