セキュリティインシデント解析支援を目的とした悪性Webサイト発見システムの提案

三須 剛史[1]　佐藤 順子[2]　花田 真樹[2]　山口 崇志[2]　布広 永示[2]

概要：近年，ドライブ・バイ・ダウンロード攻撃やフィッシング攻撃など，改ざん Web サイト及び偽装 Web サイト（悪性サイト）を経由した Web 媒介型攻撃による被害が増加している．サイバー攻撃の解析には，従来より実施されている通信ログ解析に加え， JavaScript や HTML などのソースコードやマルウェアの解析などが必要であり，これら膨大なデータ量を効率的に解析し，解析結果を統合するための計算機による支援が必要である．

本研究では，検索エンジンを用いたキーワード検索で悪性サイトの候補を抽出し，URL スキャンの結果から悪性サイト内のソースコード解析やマルウェアの動的解析を行い，解析結果の提示と悪性サイトに含まれる悪意のあるコードのシグネチャ生成を行うシステムを提案する．本稿では，キーワード検索により抽出された悪性サイト候補と URL スキャンの結果を報告する．

キーワード：MWS，悪性 Web サイト，ドライブ・バイ・ダウンロード，フィッシング

Proposal of malignant Website discovery system for security incident analysis support

Takeshi Misu[1]　Junko Sato[2]　Masaki Hanada[2]　Takashi Yamaguchi[2]　Eiji Nunohiro[2]

Abstract: The damage caused by Web-based attacks (e.g., drive-by download attacks and phishing attacks) using defacing Websites and spoofed Websites has increased. In addition to the conventional communication traffic log analysis, the analysis of malware and source code (e.g., Javascript and HTML) is needed in order to analyze cyber attacks. To proceed with the analysis of the big data efficiently, computer support is needed. In this research, we propose the system to support the analysis of security incident. In our proposed system, the candidate list of malignant sites is searched by the keyword search using the search engine. Next the candidate list of malignant sites is scanned using the URL scan site and the dynamic analysis of malware and source code is conducted. The signature code of malignant sites is created by the dynamic analysis. In this paper, we show the candidate list of malignant sites which is searched by the keyword search and the results of the URL scan of the list.

Keywords: MWS，Malignant Website，Drive-By-Download，Phishing

1. はじめに

近年，マルウェアの感染方法として，個人や企業及び自治体の Web サイト（以下，正規サイトと呼ぶ）を媒介とした攻撃が増加している [1]．例えば正規サイトのコンテンツを改ざん（以下，改ざんサイトと呼ぶ）し，マルウェア配布先の Web サイトに誘導するドライブ・バイ・ダウンロード攻撃による被害が発生している．図 1 に示すように，ドライブ・バイ・ダウンロード攻撃はマルウェア配布サイトの発見を困難にする目的で，リダイレクトコードを正規サ

[1] 東京情報大学大学院
Tokyo University of Information Sciences，Graduate School of Informatics
[2] 東京情報大学 総合情報学科
Tokyo University of Information Sciences，Department of Informatics

イトに埋め込み，多段でリダイレクトする手法が用いられる．また，オンラインバンクやインターネット通信業者，配送サービス業者を装い電子メールを送り，ユーザに偽の Web サイト（以下，偽サイトと呼ぶ）の URL にアクセスさせて，クレジットカード番号，アカウント情報（ユーザ ID，パスワードなど）といった重要な個人情報を盗み出すフィッシング攻撃による被害が発生している．図 2 に示すように，フィッシング攻撃は，偽の Web サイトのコンテンツを本物の Web サイトのコンテンツとほとんど区別がつかないように偽造し，URL の文字列も実在する Web サイトの URL に非常に類似した文字列にするなどその手口は日々巧妙化している．これらの攻撃の特徴として Web サイトを媒介として攻撃が行われるという点，改ざんサイトや偽サイトの URL は短期間で変化し，かつ Web 空間に存在する URL の数は膨大であるため特定は困難であるという点が挙げられる．

Web サイトを媒介としたこれら攻撃の解析には通信ログ解析に加え，Web サイトのコンテンツ（HTML，Javascript，テキスト）の解析やマルウェア本体の解析などが必要であり，解析における課題として，Web 空間に存在する膨大な Web データから，改ざんサイト，偽サイト，マルウェア配布サイト等の悪性サイトを効率的に発見することと，各解析の過程で得られた別々の情報を統合し提供するということが挙げられる．これにより，攻撃の被害にあった場合でも迅速な対策を講じることができると考えられる．

そこで本研究では，セキュリティインシデント解析の支援を目的として，悪性サイト発見・解析システム（以下，提案システムと呼ぶ）を提案する．提案システムは検索部，比較部，判定部，解析部，可視化部からなる．検索部では検索エンジンを用いたキーワード検索で悪性サイトの候補（以下，悪性サイト候補と呼ぶ）を抽出する．比較部では，検索部で取得した悪性サイト候補のソースコードと攻撃者が正規サイトを攻撃する際に埋め込むリダイレクトコードなどの攻撃コード（以下，シグネチャと呼ぶ）の比較を行う．判定部では検索部で取得した悪性サイト候補をオンラインの URL スキャンサイトを用いて判定を行う．解析部では仮想環境を用いて，判定部で悪性サイトと判定された Web サイトを解析する．なお，このシステムを動作させる過程で得たデータはデータベースへ保存される．可視化部では各部で収集したデータをユーザからの操作要求を元に可視化を行う．

本稿では，まず **2** 章で本研究の関連研究について述べる．**3** 章では，提案システムについて述べる．**4** 章で，提案システムの実験結果について述べる．**5** 章で，まとめと今後の課題について述べる．

2. 関連研究

本章では，悪性サイトの検出・解析手法に関する関連研究について述べる．

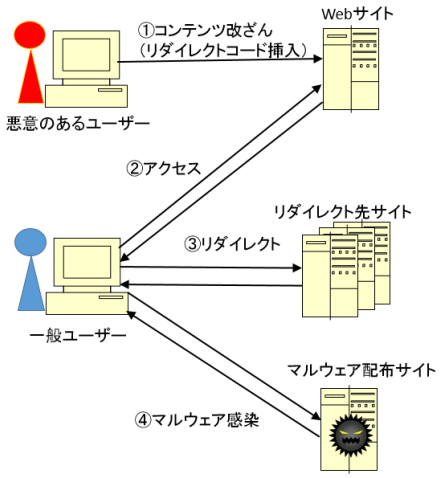

図 1 ドライブ・バイ・ダウンロード攻撃

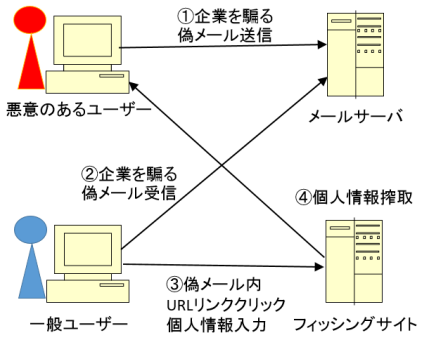

図 2 フィッシング攻撃

セキュリティ分野においては，悪性サイトの検出・解析について様々な検討が行われている [2][3][4]．[2] では，URL BlackList の構築を目的として，Web 空間に存在する未知の悪性 URL を既存の悪性 URL を用いて効率的に抽出する手法を提案している．既存の悪性 URL を起点としてクローリングを行い，収集した未知の悪性 URL を文字列長やパス文字列の平均トークン長などの特徴量を用いて既存の悪性 URL との類似度を比較し，Web クライアント型ハニーポットなどの検証ツールを用いて検証を行っている．本研究は，サイト内コンテンツ中身（ソースコード）とシグネチャでも悪性 URL の比較を行っているという点で異なる．

[3] では，DBD 攻撃の脅威把握と効果的な対策導出を目的とした攻撃対策フレームワークを提案している．複数のユーザ端末にブラウザ組み込み型センサを導入し，大規模なユーザの挙動を収集・分析し異常検知を行うことで，悪性サイトの検知を行っている．本研究では，複数のユーザ端末にセンサを導入せずに，システム単体で悪性サイトを検知する点で異なる．

[4] では，改ざんサイトの検知を目的として，検索エンジンを用いて改ざんサイト内の不正スクリプトを検出・収

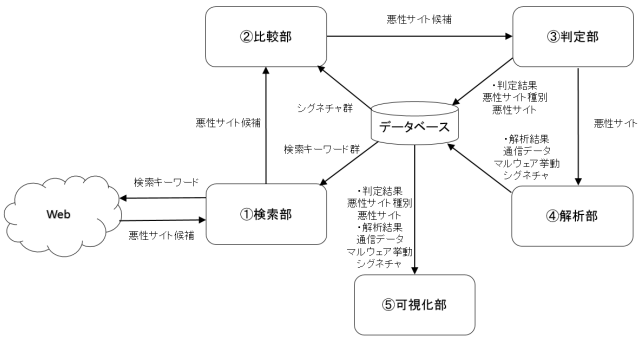

図 3 提案システム概要図

集する手法を通して，Web プロキシにおいて検知を行うシステムを提案している．検索エンジンのキーワードにはスクリプトの既存書式に既存の悪性 URL のドメインをつなぎあわせたものを用いているため，既存の悪性 URL のドメインを用いていない改ざんサイトについては検出することができない．本研究とは，検索エンジンのキーワードに既存の悪性 URL ドメインの特徴だけではなく，悪性 URL の候補を抽出するという点で異なる．

3. 提案システムについて

本章では，提案システム全体の概要について述べる．次に，提案システムにおける各部の説明と動作について述べる．最後に，提案システムの運用により期待される効果について述べる．

3.1 提案システムの構成

提案システムの概要を図 3 に示す．提案システムの 5 つの構成を以下に示す．

- 検索部

  改ざんサイトや偽サイトの特徴をキーワードとして，検索エンジンを用いてキーワード検索を行い，悪性悪性サイト候補を抽出する．具体的には，悪性サイトの多いトップレベルドメインでキーワード検索を行う方法や，改ざんされやすい脆弱性を持つサイトとして CMS を用いて書かれているサイトでキーワード検索を行うといった方法がある．キーワード検索で用いるキーワードはユーザが事前に指定してデータベースに格納することもできるが，以下で述べる判定部で，悪性サイトと判定された URL に含まれる文字列の特徴を自動抽出し，検索キーワードとして悪性サイト候補を抽出して Web 空間の絞り込みの精度を向上させる．なお，検索には Google Custom Search API[5] を用いる．次に，抽出した悪性悪性サイト候補内のリンク先を Web クローラを用いて取得する．これは，サイト内のリンクはそのサイトの内容に類似したサイトである可能性が高いという考えで行っている．検索部で得られた悪性サイト候補の URL をデータベースへ保存する．
- 比較部

  データベースから，悪性サイト候補の URL を受け取り，悪性サイト候補の URL に対してコンテンツを取得する．なお，コンテンツの取得に curl コマンドを用いた．そして，取得した悪性サイト候補のコンテンツと攻撃者が Web サイトを改ざんする際に埋め込むコード（シグネチャ）の比較を行う．比較で用いるシグネチャは事前にデータベース内に格納するが，本提案システムを動作させていく過程において，解析部の動的解析で得られた通信データ (HTML，テキスト，Javascript) からシグネチャを自動生成し，生成したシグネチャを比較部のシグネチャとして追加・更新することで，比較精度の向上を図る．. 悪性サイト候補のコンテンツ内にシグネチャが含まれていた場合はソースコードとシグネチャの対応付けを行いデータベースへ保存する．
- 判定部

  データベースから，悪性サイト候補の URL を受け取り，オンライン URL スキャンサイトを用いて悪性サイトであるか判定を行う．オンライン URL スキャンサイトには Virustotal[6] を用いる．VirusTotal とは疑わしいファイルやウェブサイトのマルウェア検査を行うウェブサービスである．複数の既存アンチウイルス製品を用いて解析を行い，解析結果をユーザに提示する．悪性サイト候補の URL が Virustotal で悪性サイトと判定された場合は悪性サイトの URL と解析結果の対応付けを行いデータベースへ保存する．判定部の解析結果は，そのサイトがどのような種類の悪性サイトか，いくつのアンチウイルス製品にヒットしたか，という簡易的な情報のため，悪性サイトと判定された場合，より詳細な情報を取得するために解析部に悪性サイトと判定された URL を受け渡す．
- 解析部

  データベースから，Virustotal で悪性サイトだと判定された URL に仮想環境上でアクセスし動的解析を行う．悪性サイトアクセス時の通信データ（javascript，HTML，テキスト，ヘッダー，リダイレクト回数，IP アドレス，通信間隔等）やマルウェアの挙動（レジストリの書き換え，呼び出す API 等）を動的解析する．動的解析には高対話型のハニーポットとサンドボックスを用いる．動的解析の結果はデータベースへ保存する．
- 可視化部

  データベースに保存された検索部，比較部，判定部，解析部のデータを統合し，ユーザの要求に応じて可視化する．検索キーワードとシグネチャの関係性を図で表示する，各シグネチャのヒット数の時系列推移

を線グラフで表示する，改ざんサイトのソースコードとリダイレクト先及びマルウェアの挙動の関係性を図で表示するなどが挙げられる．

3.2 提案システムの効果

提案システムの運用における期待される効果は大きく 3 つある．

1 つ目は解析部で得られた動的解析結果からシグネチャを自動で生成するということである．解析部で解析を行う URL は，Web 空間から検索を行ったものであり，比較部及び判定部を通して悪性だと判断された URL はすぐに解析を行い，その結果から攻撃コードのシグネチャを自動生成するため鮮度が高いため，日々刻々と変化するサイバー攻撃の手法を把握する支援になることが期待される．

2 つ目は判定部で得られた解析結果を基に，悪性サイトと判定された URL を基に，検索キーワードを自動生成するということである．自動抽出した検索キーワードを基に本システムを運用し，悪性サイトの抽出数などで危険度を付与し，危険度が上位の悪性サイトの URL を Black List に登録することで，サイバー攻撃の対策支援になることが期待される．

3 つ目は提案システムの運用によって得られた各部での情報を統合してユーザに提供するということである．セキュリティインシデントの分析を手動で行う場合，ユーザは各解析過程で得られた情報を一旦統合してから俯瞰的にみる必要があるが，提案システムではユーザは必要な情報を指定するだけで，統合された情報を可視化することができるため，セキュリティインシデント分析の支援になることが期待される．

4. システム運用実験

本章では，2016 年 8 月 2 日〜8 月 9 日の間システムを運用した結果について述べる．今回のシステム運用実験では，システムの各部における動作実験として以下の項目を明らかにする．

- 検索部におけるキーワード別悪性サイト候補抽出数
- 比較部におけるシグネチャの種類
- 判定部における検索部で抽出した悪性サイト候補と検索キーワード別の判定結果

システム運用実験の結果，いくつかの悪性サイトを抽出可能であるということが明らかになり，トップレベルドメインと悪性サイト数のおおまかな傾向をみることができた．なお，解析部と可視化部については現在実装中であるため，今後システム運用実験を行う予定である．

4.1 検索部における悪性サイト候補抽出結果

今回の実験で用いた Google Custom Search API の検索キーワードを表 1 に示す．今回の運用実験において検索

表 1 検索キーワード一覧

site:.xyz inurl:index.php
site:.ru inurl:index.php
site:.info inurl:index.php
site:.com inurl:index.php
site:.cn inurl:index.php
site:.hk inurl:index.php
site:.cm inurl:index.php
site:.jp inurl:index.php
site:.org inurl:index.php
site:.ro inurl:index.php

表 2 検索結果一覧

検索キーワード	検索ヒット数
site:.xyz inurl:index.php	94 件
site:.ru inurl:index.php	175 件
site:.info inurl:index.php	64 件
site:.com inurl:index.php	2691 件
site:.cn inurl:index.php	1082 件
site:.hk inurl:index.php	659 件
site:.cm inurl:index.php	259 件
site:.jp inurl:index.php	172 件
site:.org inurl:index.php	1108 件
site:.ro inurl:index.php	607 件
合計	6911 件

方法は検索演算子を用いた．検索演算子にはいくつか種類があるが，今回の実験では「site:」と「inurl:」を用いた．「site:」はドメインを指定して検索を行う，「inurl:」は URL に含まれる文字列を検索するという意味である．検索キーワードにドメイン指定を用いたのは，サイバー犯罪者が悪意のあるサイトのドメインを登録する際に，規制が少なく登録が容易であり価格の安い場所を選択する傾向がみられ，悪性サイトの温床となるドメインが存在するためである．「.jp」を除くドメインについては危険とされているドメインを用いた [7]．検索演算子「inurl:」に「index.php」を用いたのは，php には数多くの脆弱性があるという点と，多くの Web サイトで利用されているという点から，攻撃者に狙われやすいのではないかという理由である．また，攻撃者が Web サイトを改ざんする際にサイト構造で階層の深いディレクトリにあるページを改ざんするよりも，多くのユーザが閲覧する可能性の高いトップページを改ざんするのではないかという理由もある．

表 1 の検索キーワードを用いて検索を行ったヒット数を表 2 に示す．総取得 URL 数は 8086 件であった．その内検索キーワードで用いたドメインが含まれる URL は 6911 件である．ドメイン別に見ると「.com」のドメインが最も多く，「.info」のドメインが最も少ないことがわかる．

なお，1175 件の URL は検索キーワードが含まれず，その多くは

- IRC，FTP プロトコル

- URL が IP アドレス
- localhost ドメイン

であった．これは，Web クローラで取得した URL は悪性サイト候補のソースコードに含まれる全てのリンク先を抽出するため，検索キーワードが含まれない URL が抽出されたと考えられ，今後改善の必要がある．

なお，検索方法は表 1 のキーワードだけでなくユーザが任意に指定可能であるが，3 章で述べたように，今後は判定部で悪性サイトと判定されたコンテンツのソースコードと URL の特徴を自動抽出し，検索部の検索キーワードとすることで悪性サイト候補の抽出も行う．自動抽出した検索キーワードには頻出度で重み付けを行い，優先度の高い検索キーワードから検索を行う．これにより，悪性サイト候補をより絞ることが可能となり，Web 空間を効率よく探索できると考える．この他にも，ソースコード内に記述されている言語と，トップレベルドメインが違う場合，悪性サイトの可能性があるとして，検索を行うなどが考えられる．

4.2 比較部における悪性サイト候補とシグネチャの比較結果

検索部で抽出した悪性サイト候補の URL を比較部にて比較を行ったが，本実験で使用した 33 種類のシグネチャの内，ヒットしたのは 4 種類であった．その内訳は「eval」「document.write」「escape」「unescape」であった．これらは正規サイトで用いる Javascript にも使用されることもあるため，今回の実験では比較部による悪性サイト候補の絞り込みは困難であった．

3 章で述べたように，今後は解析部の動的解析で得られた通信データ (HTML，テキスト，Javascript) からシグネチャを自動で生成し，生成したシグネチャは比較部のシグネチャとして活用することで，比較精度の向上を図る予定である．

4.3 判定部における悪性サイト候補の判定結果

本提案システムで抽出した悪性サイト候補の URL を Virustotal で URL スキャンを行った．判定結果のドメイン別分類一覧を表 3 に示す．判定数とは 1 つ以上のアンチウイルス製品に判定された場合を 1 件とし，スキャン件数は，1 つの URL が複数の既存アンチウイルス製品に判定された場合の総数である．悪性サイト候補 6911 件中 23 件が悪性サイトであるという判定になった．ドメイン別に見ると「.ru」のドメインが 11 件と最も多く，「.xyz，.info，.cn」のドメインが 1 件と最も少ないことがわかる．一方でスキャン数をみると「.ru」のドメインが 39 件と最も多く，次いで「.xyz，.com」のドメインが 4 件という結果になった．判定数が最も少なかった「.xyz」が上位に上がったのは，ドメインが「.xyz」の悪性サイト 1 件に対して，4 つのアンチウイルス製品に悪性サイトだと判定されたということであり，同じスキャン数の「.com」と比べると，より悪性サイトの可能性が高いと言える．なお，判定部で得られた悪性サイトの URL を解析部に受け渡し，仮想環境上で動的解析を行う．

表 3 ドメイン別分類一覧

トップレベルドメイン	判定数	スキャン数
.xyz	1 件	4 件
.ru	11 件	39 件
.info	1 件	1 件
.com	4 件	4 件
.cn	1 件	3 件
.cm	3 件	3 件
.org	2 件	2 件
合計	23 件	56 件

5. まとめと今後の課題

本研究では，セキュリティインシデント解析の支援を目的として，Web 空間から悪性サイト候補の URL を収集し判定・解析・可視化するシステムを構築した．システム運用実験の結果，いくつかの悪性サイトを Web 空間から抽出した．現在実装している項目は以下の通りである．

- 悪性サイトのソースコードからシグネチャを自動生成 (解析部)
- 悪性サイトの URL から検索キーワードを自動生成 (判定部)
- Web 媒介型攻撃を対象とした仮想解析環境 (解析部)
- 提案システムで収集したデータの可視化 (可視化部)

今後は解析部と可視化部のシステム運用実験を行うとともに，新たな検索キーワードとシグネチャを用いて提案システムを運用する予定である．

謝辞 本研究を進めるにあたって，株式会社日立システムズ関係者各位に助言と協力を頂きました．ここに深く感謝します．

参考文献

[1] 独立行政法人情報処理推進機構，2014 年度情報セキュリティ事象被害状況調査，http://www.ipa.go.jp/security/fy26/reports/isec-survey/

[2] 孫 博，秋山 満昭，八木 毅 他，“広大な Web 空間を対象とした悪性 URL の検索技術”，電子情報通信学会技術研究報告，ICSS，情報通信システムセキュリティ，vol.114，pp61-66(2014)．

[3] 笠間 貴弘，井上 大介，衛藤 将史，“ドライブ・バイ・ダウンロード攻撃対策フレームワークの提案”，コンピュータセキュリティシンポジウム 2011 論文集，pp.780-785(2011)．

[4] 田村 佑輔，甲斐 俊文，佐々木 良一，“ユーザ標的型 Web サイト改ざんに対する検索エンジンを用いた検知手法の提案”，情報処理学会論文誌，vol. 51，pp. 191-198(2010)．

[5] Google Custom Search API，https://developers.google.com/custom-search/?hl=ja

[6] Virustotal online service，https://www.virustotal.com/ja/

[7] McAfee 危険な Web サイトの世界分布，https://promos.mcafee.com/ja-JP/PDF/Mapping_Mal_Web_Summary.pdf