卒業論文　　2012年度（平成24年度）

効率的な悪性プログラム収集システムの設計と実装

慶應義塾大学 環境情報学部

氏名：吉原大道

担当教員

慶應義塾大学 環境情報学部

村井 純

徳田 英幸

楠本 博之

中村 修

高汐 一紀

Rodney D. Van Meter III

植原 啓介

三次 仁

中澤 仁

武田 圭史

平成25 年1 月22 日

卒業論文要旨 - 2012 年度 (平成 24 年度)

効率的な悪性プログラム収集システムの設計と実装

近年，ウイルスやワームなどの悪意のあるソフトウェア（以下，マルウェア）による脅威が顕在化している．現在では，様々な種類のマルウェアが日々出現し，多様な感染活動が存在している．このため，多様化するマルウェアに対して効果的な研究や対策をとるには，マルウェアによる攻撃の傾向および特徴を得るとともに，より多くのマルウェアを効率的に収集する必要がある．本論文では，能動型のハニーポットを用いた収集環境を構築し，これまでのデータからマルウェア配布サイトの特徴を推測することで新規のマルウェアを効率的に収集することを目標とする．これによって，多様なマルウェアを利用した効果的な研究活動を行うことができると期待される．それを実現するために，短時間でより多くの，マルウェアを配布する悪意のある Web サイトにアクセスするための手法を提案した．そして，提案した手法の有効性を実証するため，マルウェアの検体を収集するシステムを実装した．本システムでは，提案手法を用いて自動で Web クローリングを行い，収集した大量の URL から悪意のある Web サイトのものである URL を探し出す．そして，その URL に優先的にアクセスすることで効率的なマルウェア検体の収集を可能にする．また，実際にそのシステムを使用し，検証を試みた．その結果，本論文にて提案した手法を用いることで大量の URL の中から悪意のある Web サイトのものである URL を見つけることができた．これにより，効率的にマルウェア検体を収集することができるということを確認した．本論文の成果により，マルウェアの収集を効率的に行うことができ，増加を続けるマルウェアによる脅威に対して効果的な対応をとっていくことができると期待される．

キーワード:
1. マルウェア, 2. ハニーポット, 3. セキュリティ, 4. インターネット

慶應義塾大学 環境情報学部

吉原 大道

Abstract of Bachelor's Thesis - Academic Year 2012

Design and Implementation of Efficient Malware Collecting Systems

These days, malicious programs (from now, "malware") such as computer viruses and worms have become a distinct threat. These days, all kinds of malware emerge everyday and so there are various infection activities. For this reason, we must determine the tendencies and features of malware, and collect as many malware as possible in an efficient way in order to efficiently study and take measures of the real-time-changing malware. The purpose of this thesis is to make a collecting environment using an active honeypot, and collect new malware in an efficient way by speculating the features of websites which distribute malware from existing data. It is expected that an efficient study which use various malware will be made. To achieve this, I suggest a method to access to many malicious websites that distribute malware within a short time. To prove the effectiveness of this method, I will use it and do an automatic web crawling. This implements a system which collects samples of malware, by finding and preferentially accessing to the URLs of malicious websites from among a massive quantity of URLs. Also, I used this system and did a test. As a result, by using this method, I was able to find URLs of malicious websites out of many URLs, and confirmed that it is able to effectively collect malware samples. It is expected that from the result of this thesis, it will become able to efficiently collect malware and make an effective correspondence against the threat of increasing malware.

Keywords :
1. Malware, 2.Honeypot, 3.Internet Security, 4. Internet

Keio University, Faculty of Environment and Information Studies

Daido Yoshihara

目 次

図 目 次

表 目 次

第1章　序論

本章では研究の背景として，マルウェアの脅威と対策やそれを取り巻く環境の現状について述べる．はじめに，マルウェアの増加や多様化によってその脅威が増大しており，解析・検知を迅速に行う必要があるということについて述べる．次に，こうした対策技術を向上させるためにはマルウェア検体が必要であるということについて述べ，そのためには効率的なマルウェア検体の収集環境が必要であるという現状を述べる．そして，マルウェアの検知・解析を迅速に行う研究者を支援するため，効率的にマルウェア検体を収集できる手法を確立するという目的を明らかにする．最後に本論文中で用いる用語を定義し，本論文の構成を記す．

1.1　マルウェアの脅威とそれを取り巻く環境の現状

近年，マルウェアの種類や感染活動が多様化している．ウイルス対策ソフトウェアベンダであるMcAfee社 [1] によるMcAfee脅威レポート2012年第1四半期 [2] によると，2012年に入り，McAfee Labsでは7,500万件以上の新しいマルウェアのサンプルを検出している．これまで鎮静化していたパスワード盗用型トロイの木馬やZeroAccessルートキットや署名付きマルウェアといった新たな脅威に加え，新しい時系列のルートキット，携帯端末を狙うマルウェアなど様々な種類のマルウェアによる脅威が存在するとされている．さらに，2000年代中盤より，マルウェア感染の原因の大半は悪意のあるWebサイト閲覧などのインターネット経由での感染であるとトレンドマイクロ社 [3] は述べている．このため，多様化するマルウェアに対して効果的な研究や対策を行うには，攻撃の傾向および特徴を得るとともに，より多くのマルウェアを効率的に収集する必要がある．マルウェアを収集する方法として，ハニーポットと呼ばれる囮のシステムを用いる方法が存在する．ハニーポットには大きく分けて2つの種類に分けることができる．1つ目は，脆弱性を突いて行われるサーバへの攻撃を検出・解析するために受動的に攻撃を待ち受ける種類のものである．そして2つ目は，悪意のあるWebサイトにアクセスすることで能動的に攻撃を捕捉する種類のものである．収集する攻撃の種類によってどちらのハニーポットも必要であるが，近年は受動的な攻撃の情報を収集するクライアント型ハニーポットが多く利用されている．このような能動的に悪意のあるWebサイトにアクセスしてマルウェア検体を収集するハニーポットでは，短時間でより多くのWebサイトを巡回する機能が必要であり，その手法が求められている．

1.2 本論文の目的

本論文の目的は，効率的に悪意のある Web サイトを巡回することができるシステムを用いることで，効率的にマルウェア検体を収集することである．これにより，マルウェアの解析や検知を行う研究者にマルウェア検体の情報を効率的に提供することができ，より効果的な対策を迅速にとることができるようになると期待される．

1.3 本論文中の用語の定義

本論文では，悪意のある Web サイトおよび種（Seed）となる URL という用語を用いる．ここでは，これらの用語についての定義を行う．本論文における悪意のある Web サイトとは，Drive-by download attack を行う Web サイトのことであると定義する．Drive-by download attack とは，ユーザが悪意のある Web サイトにアクセスした際に，マルウェアに感染させる攻撃である．また，本論文における種（Seed）となる URL とは，悪意のある Web サイトを巡回する際の基準点となる URL であると定義する．本論文では,「悪意のある Web サイトは同ドメイン内の異なるパスに存在する」と述べる秋山らの研究 [4] に基づき，悪意のある Web サイトの URL を種（SeedURL）となる URL とし，Web クローリングを行う際の基準点としている．また，本論文で行った検証実験の結果に伴い，FP・FN 及び TP・TN という略語を用いる．FP とは，False Positive の略称であり，正常な Web サイトを悪意のある Web サイトであると判定してしまう「誤検知」と定義する．FN とは，False Negative の略称であり，悪意のある Web サイトを正常な Web サイトであると判定してしまう「検出漏れ」と定義する．TP とは，True Positive の略称であり，正常な Web サイトを正常な Web サイトであると正しく判定する「正検出」と定義する．TN とは，True Negative の略称であり，悪意のある Web サイトを悪意のある Web サイトであると正しく判定する「真陽性」と定義する．

1.4 本論文の構成

本論文は全 8 章から構成される．第 2 章では，マルウェアによる脅威とそれを取り巻く環境の現状について述べる．第 3 章では，第 2 章で述べた課題に取り組む関連研究を紹介する．第 4 章では，効率的にマルウェアを収集するための手法を提案する．第 5 章では，第 4 章で述べた手法に基づき構築した，効率的にマルウェア検体を収集するシステムの実装について述べる．第 6 章では，実装したシステムを使用して行ったいくつかの実験と，その結果について述べる．第 7 章では，第 6 章での結果を基に，システムについて様々な面から評価を行い，考察を与える．最後に，第 8 章で本論文の結論と今後の展望を述べる．

第2章　マルウェアの現状

本章では，現在におけるインターネット上でのマルウェアの感染活動の状況及び多様化する種類について述べる．また，そうしたマルウェアへの対策の現状についても述べる．そしてその中で，本研究がマルウェア対策において果たす役割を示す．

2.1　マルウェアとは

マルウェアとは，不正かつ有害な動作を行う意図で作成された悪意のあるソフトウェアを意味する Malicious Software を短縮した造語である．後述する，コンピュータウイルスやトロイの木馬，バックドアなどの不正プログラムを総称する単語として用いられている．

2.1.1　マルウェアの種類と脅威の事例

ここでは，現在どのような種類のマルウェアが存在するのかをまとめる．また，それぞれのマルウェアが実際にもたらす脅威について，近年報告された事件などを紹介することで述べる．

- **コンピュータウイルス**
  コンピュータウイルスとは，第三者のプログラムやデータベースに対して意図的に何らかの被害を及ぼすように作られたプログラムである．また，自己伝染機能・潜伏機能・発病機能のうちの一つ以上の機能を有するものであるとされている．広義ではコンピュータに被害をもたらす不正なプログラムの一種であり，以下で紹介するマルウェアの総称でもある．

- **ワーム**
  ワームとは，自身を複製することで他のシステムに拡散する性質を持つ独立したプログラムである．宿主となるファイルを必要としないという点において狭義のコンピュータウイルスとは区別される．2009 年には Stuxnet と呼ばれるワームによってイランの核施設が妨害されるという事件が起きた．また，2012 年にはソーシャルネットワークサービスである Facebook[5] のログイン情報が 4 万 5000 件以上盗まれていたことが Seculert 社 [6] によって報告されている．この事件では 2010 年に発見された Ramnit と呼ばれるマルウェアの亜種である Ramnit.C と呼ばれるワームが用いられた．

- **トロイの木馬**
  トロイの木馬とは，正常なソフトウェアを装うことでユーザに自身をダウンロードさせ，実行させるソフトウェアである．バックドア型やパスワード窃盗型，ダウンローダ型など様々な種類に分類することができる．自己増殖機能がないという点において，狭義の意味でのコンピュータウイルスとは区別されている．2011 年には標的型メール攻撃により，メールを開いた衆議院のサーバ及び議員の端末がパスワード窃盗型のトロイの木馬に感染した．この事件では，議員全員の ID・パスワードが外部に流出しメールなども外部から閲覧されていた．また，2012 年にも同様の手口で宇宙航空研究開発機構（JAXA）[7] の職員の端末がバックドア型のトロイの木馬に感染し，情報が外部に流出したという事件も起きている．このようにトロイの木馬の感染による様々な事件が大きな問題となっている．

- **スパイウェア**
  スパイウェアとは，ユーザが認識しないうちにバックグラウンドにて動作し，ユーザのブラウジング履歴や個人情報などを収集してマーケティング会社など特定の相手に送信するソフトウェアである．2005 年にインターネットバンキング利用者のパスワードなどを盗み，不正な振り込みを行った人物が逮捕される事件が起きた．この事件のように，個人規模での感染事例が非常に多い．

- **バックドア**
  バックドアとは，他人に知られることなくコンピュータ内に設けられた通信接続の機能をもつソフトウェアである．その目的は，ID やパスワードを使って通信を制限したり使用権を確認したりするコンピュータの機能を，無許可で利用するためである．2011 年の三菱重工の事例では，三菱重工 11 拠点の 83 台の端末にバックドア及び先述したスパイウェアを含む 50 種類以上のマルウェアが使用された．

- **キーロガー**
  キーロガーとはキーボードからの入力を監視して記録するソフトウェアである．複数の人間が利用するパソコンに仕掛けることでパスワードやクレジットカードの番号などを収集して特定の相手に送信するなど，悪用されることが多い．インターネットカフェに仕掛けられたキーロガーにより，ネットバンキングのパスワードやカード番号などの個人情報が盗まれるといった被害が多くみられる．

- **アドウェア**
  アドウェアとは，広告を目的としたソフトウェアである．基本的には無害であるが，中にはユーザに告知せず情報を収集するマルウェアであるものも存在する．アドウェアは大きく 2 つに分類される．1 つ目は，ブラウザを使用していないにも関わらずポップアップ広告を表示させる機能を持つポップアップ広告型である．2 つ目はリンク先を書きかえることで Web サイトの閲覧中に別の広告サイトにページが切り替えるリンク乗っ取り型である．ウイルス対策ソフトウェアベンダであるトレンドマイクロ社 [3] は，近年のスマートフォンの急激な普及により，Android OS を搭載したモバイル端末がアドウェアの被害を受ける危険性について注意を促している．

2.1.2 マルウェアによる脅威の現状

ウイルス対策ソフトウェアベンダである McAfee 社 [1] が掲示する，McAfee 脅威レポート 2012 年第 1 四半期 [2] によると，2011 年の終わりには多くの地域でマルウェアの脅威の減少が確認されている．しかし，現在は対極の状況であり，PC を攻撃するマルウェアの数も近年の中で最も多い状況にあることが示されている．2012 年に入り，McAfee Labs では累計 7,500 万件以上のマルウェアのサンプルが検出されている．この脅威レポートによると，McAfee のデータベースにはすでに 8,300 万件のマルウェアが登録されており，2012 年の第 2 四半期もしくは第 3 四半期の間に 1 億件に達することは確実であると考察されている．さらに，これまで鎮静化していたパスワード盗用型トロイの木馬に加え，ZeroAccess ルートキットや署名付きマルウェアなどの新たな脅威や，新しい時系列のルートキット，携帯端末を狙うマルウェアといった様々な種類のマルウェアの脅威が存在すると，McAfee 脅威レポート 2012 年第 1 四半期は述べている．このように，多種多様なマルウェアがインターネット上に存在し，その数を増やし続けている．このことから，2012 年以降もマルウェアによる脅威は増加していくということが考えられる．

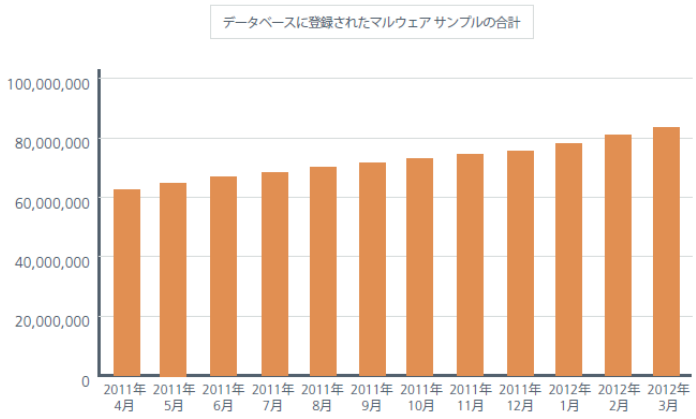


図 2.1: McAfee 脅威レポート 2012 年第 1 四半期による McAfee Labs のデータベースに登録されたマルウェアサンプルの合計

2.1.3 マルウェアの感染経路

マルウェア感染の経路は種類によって様々である．以下にマルウェアの感染経路の種類について述べる．

- **Web サイトの脆弱性を利用した受動型攻撃による感染**
  受動型攻撃とは，ユーザが Web サイトにアクセスするなどの行動を起こした際に，攻撃者から悪意のあるデータを送信され任意のコードを実行されてしまう攻撃である．トレンドマイクロ社 [3] は 2000 年代中盤より，マルウェア感染の原因の大半は悪意のある Web サイトを閲覧した際に感染するといったような，インターネットを経由したものであると発表している．そうしたインターネット経由でのマルウェア感染に，Drive-by download attack が存在する．Drive-by download attack とは，ユーザが悪意のある Web サイトを閲覧した際に，攻撃者がユーザに気付かれないようにマルウェアなどのソフトウェアをダウンロードさせる攻撃である．この攻撃は，攻撃者が Web ブラウザ本体の脆弱性や Web ブラウザのプラグインの脆弱性などを利用することで生じる．攻撃者は，難読化した JavaScript や HTML の iframe タグを利用してユーザを悪意のある Web サイトに誘導する．ESET 社 [8] 提供による月刊マルウェアランキング [9] の図 2.2 によると，2012 年の 9 月に日本で流行したマルウェアのトップ 10 のうち，3 割が難読化された JavaScript が含まれていることが分かっている．

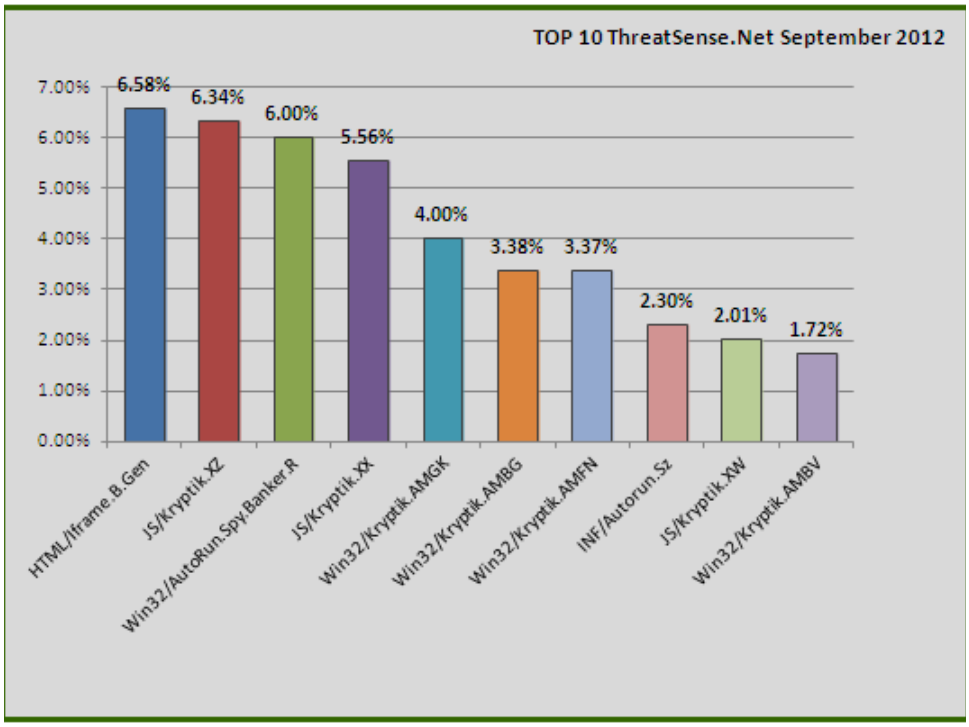


図 2.2: ESET 製品 ThreatSense.Net によるマルウェアランキングトップ 10（2012 年 9 月）

また，一般の Web サイトに対して SQL インジェクション攻撃を行い，その Web サイトに他の悪意のある Web サイトに誘導するコードを設置する方法も存在する．2009 年末から 2010 年初頭にかけての gumblar 攻撃では，改ざんされた正規のサイトにアクセスした際に悪意のある Web サイトに転送され，気付かないうちにマルウェアに感染してしまうユーザが多く存在した．
本論文では，このような受動型攻撃に注目し，マルウェアを収集している．

- **ネットワークサービスの脆弱性を利用した能動型攻撃による感染**
  能動型攻撃とは，攻撃者から悪意のあるデータを送信され任意のコードを実行されてしまう攻撃である．ユーザが特に行動を起こさない場合でも攻撃者が能動的に行動を起こすことで，悪意のあるデータを送信され任意のコードを実行されてしまうという点で受動型攻撃とは大きく異なる．
  そのような能動型攻撃の情報を収集するハニーポットとして Nepenthes[10] というツールが存在する．本研究の事前研究として Nepenthes を利用し，情報の収集を行っている．詳細は第 2.3.3 項にて述べるが，結論として，能動型攻撃の情報を収集するハニーポットよりも，受動型攻撃の情報を収集するハニーポットの方が効率よくマルウェア検体を収集できるということが分かっている．そのため，一般的なユーザがインターネットを利用する際には，能動型攻撃よりも受動型攻撃の方が遭遇する可能性が高いということが言える．

- **標的型メール経由による感染**
  電子メールの添付ファイルを開いた際，もしくはメール本文に記載されているリンクにアクセスした際にマルウェアに感染するという経路が存在する．先述したように衆議院の議員や宇宙航空研究開発機構（JAXA）の職員が，マルウェアの添付されたメールを閲覧した際にマルウェアに感染した．これにより機密情報が外部に流出したという事件が，2011 年と 2012 年にそれぞれ起きている．

- **物理ドライブ経由による感染**
  USB メモリなどの移動可能な記憶装置を媒体として感染する経路が存在する．USB メモリからの感染は，外部メディア内のファイルを自動で実行するためのプログラムファイルを悪用することで起きる．この経路での感染は，インターネットに接続していない閉鎖された環境でも起こりうるものである．

このように，マルウェアには多様な特徴を持つ個体が存在し，様々な感染経路をとることでユーザを危険にさらしている．さらに，前節で述べたようにマルウェアはその数を年々増やし続けている．そのため，現存するマルウェア検体の挙動を解析し対策をとることが非常に重要であり，そうした対策をとるためにも，継続的なマルウェア検体の収集を迅速かつ効率的に行う必要がある．

2.2 マルウェアへの対策

本項では，マルウェアへの対策について述べる．現在，マルウェアの感染を防ぐにはウイルス対策ソフトウェアの導入が最も効果的かつ導入コストが低い．ウイルス対策ソフトウェアベンダでは，マルウェアを検知し解析を行うことでマルウェアへの対策を行っている．また，ウイルス対策ソフトウェアの他にもいくつかのサービスが存在する．例として，インターネットサービスプロバイダが個人もしくは小規模ユーザ向けに提供するメールゲートウェイ型のウイルスチェックサービスが存在する．他にも，企業向けに提供するルーターやファイアウォール機器にマルウェアや不正アクセスの検出及び遮断機能を持たせるサービスが存在する．

2.2.1 マルウェアの検知

ここではウイルス対策ソフトウェアベンダが行うマルウェア検知の方法について述べる．ウイルス対策ソフトウェアベンダでは，以下の手法を用いてマルウェアの検知を行っている．

パターンマッチング手法

パターンマッチング手法とはマルウェアの特徴をパターンとしてリストにし，そのパターンに合致したものをマルウェアとして検出する手法である．ウイルス対策ソフトウェアベンダは，マルウェアのパターンファイルやシグネチャファイルをマルウェア定義ファイルとして随時更新し，検査対象プログラムがマルウェアに見られる特徴的なコードを含んでいるかどうかを判定する．しかし，この手法では既存のマルウェアを検知することは可能である一方，リストにない未知のマルウェアを検知することはできない．そうした未知のマルウェアを検知する方法の 1 つとしてヒューリスティック手法が挙げられる．

ヒューリスティック手法

ヒューリスティック手法とはマルウェアの取り得る挙動をリストにし，検査対象プログラムに含まれる挙動と比較することでマルウェアを検出する手法である．ヒューリスティック手法には静的ヒューリスティック手法と，ビヘイビア方法とも呼ばれる動的ヒューリスティック手法が存在する．静的ヒューリスティック手法では，マルウェアがとるであろう処理のコードをチェックし，リスト化する．しかし，プログラム部分が暗号化されている場合，この手法ではコードを直接チェックすることができない．このように静的ヒューリスティック方法が検出を苦手とするマルウェアは動的ヒューリスティック方法にて検知を行う．例えば，暗号化型や多形態型，自己改変型のマルウェアは動的ヒューリスティック方法にて検知する．動的ヒューリスティック方法には，以下の 2 つの方法がある．1 つは，検査対象プログラムを直接実行して危険な行動を検出した時点でその動作を停止させる方法である．もう 1 つは仮想環境で検査対象プログラムを実行して危険な行動を検出する方法である．しかし，これらの方法には，プログラムを実行してしまう危険性や，マルウェアが仮想

環境を識別してしまい検出できないといった問題点も存在する．ヒューリスティック手法では，未知のマルウェアを検知できる一方で，フォールスポジティブやフォールスネガティブといった問題が存在する．フォールスポジティブとは，検査対象プログラムがマルウェアではないにもかかわらず，そのプログラムがマルウェアであると誤検知を起こしてしまう問題である．フォールスネガティブとは，検査対象プログラムがマルウェアであるにもかかわらず，そのプログラムがマルウェアでは無いとして検知漏れを起こしてしまうという問題である．

このように，検知を行うことで検知用のデータを充実させるためにも，マルウェア検体の収集を効率的に行う必要があるということを改めて述べておく．

2.2.2 マルウェアの解析

本項では，マルウェアの解析について述べる．マルウェアを解析する方法としては，動的解析と静的解析の 2 つに大きく分類することができる．

動的解析

動的解析（ブラックボックス手法）とは，実際にマルウェアを動作させ感染活動を確認することで，マルウェアの挙動やもたらす被害を明らかにする手法である．動的解析は短時間で容易に挙動を把握することができる．しかし，解析に使用するマシンが実際にマルウェアに感染してしまう危険や特定の条件下でのみ動作するマルウェアの挙動を調査することができないという欠点も存在する．

静的解析

静的解析（ホワイトボックス手法）とは，実際にマルウェアを動作させずにリバースエンジニアリングを行い，マルウェアの構造や仕様を分析する手法である．静的解析は安全な環境で完全にマルウェアの挙動を把握することができる．しかし，解析に時間がかかることや，分析を行うにはある程度の知識や経験が必要であるという欠点も存在する．

実際のマルウェア解析においては，動的解析と静的解析を組み合わせて行うことが効果的である．

2.3 マルウェアの収集

前節で述べたマルウェア対策のための検知，解析を行うためにはまずマルウェア検体を収集する必要性がある．ここでは，マルウェアの収集の方法について述べる．マルウェアの収集には主に，ハニーポットという囮手法が用いられる．

2.3.1 ハニーポット

ハニーポットとは，不正アクセスを受けることに価値を持つ囮のシステムもしくはその手法のことである．ハニーポットを設置する目的として，マルウェア検体の収集や不正アクセスの手法や傾向の解析，侵入者の攻撃目標を重要なシステムから逸らすことなどがある．従来，ハニーポットはサーバへの攻撃を検出・解析するため受動的に攻撃を待ち受けるものであった．しかし攻撃手法の変貌により，悪意のある Web サイトにアクセスすることで能動的に攻撃を捕捉するクライアント型ハニーポットが開発，利用されている．ハニーポットは，使用時にユーザが背負うリスクの大きさによって以下の 2 つに分類することができる．

高対話型ハニーポット

高対話型ハニーポットとは，実際の OS や脆弱性のあるソフトウェアを使用するハニーポットである．実環境を用いるため，多くの情報を得ることができるその反面，実際にシステムに侵入されたりマルウェアに感染するリスクも存在する．

低対話型ハニーポット

低対話型ハニーポットとは，特定の OS やアプリケーションをエミュレートして運用するハニーポットである．高対話型ハニーポットと比べて得ることのできる情報量は劣るが，安全に運用することができる．代表的なものに能動型攻撃の情報を収集するハニーポットである Nepenthes が挙げられる．
本研究を行う際に事前研究として Nepenthes を使用してマルウェア検体を収集している．第 2.3.3 項にて結果及び詳細を述べる．

また，ハニーポットは収集する情報の種類によって以下の 2 つに分類することができる．

Web サーバ型ハニーポット

Web サーバ型ハニーポットとは，Web アプリケーションの脆弱性を標的とした攻撃の情報を収集するためのハニーポットである．Web サイトをマルウェアに感染させる手法としては RFI（Remote File Inclusion）攻撃が存在する．RFI 攻撃とは，攻撃者が Web アプリケーションの脆弱性を利用することで悪意のある Web サイトに誘導し，マルウェアをダウンロードさせる攻撃である．RFI 攻撃は，近年多く観測されており，大きな脅威となっている．谷本らの研究 [11] では，Web ハニーポットを用いて効率的に攻撃元 IP や悪意のある Web サイトの URL 情報を収集するためには，多数の Web ハニーポットの運用が必要である可能性が高いことがわかっている．

Web クライアント型ハニーポット

Web クライアント型ハニーポットとは，脆弱性のあるクライアントソフトウェアを動かしながら悪意のある Web サイトにアクセスしその後のシステムの挙動を監視することで，攻撃に関する情報を収集するハニーポットである．クライアント型ハニーポットには様々なツールが存在し，北村らの研究 [12] では，5 つのクライアント型ハニーポットツールを比較している．その中で Capture-HPC が最も検出手法

が優れていると示されている．しかし，この Capture-HPC では攻撃経路の情報が得られないという課題も存在している．また，秋山らの研究 [13] では以下の 4 点がクライアント型ハニーポットに必要な事項とされている．

1. **検出精度と多様性**
   フォールスポジティブ（誤検知）とフォールスネガティブ（検知漏れ）を極力少なくし，様々な攻撃を正確に検出すべきである．
2. **多様な検体の収集**
   様々な形式の検体を収集する必要がある．
3. **効率的なパフォーマンス**
   膨大な広さの Web 空間を迅速にクローリングし，効率よく巡回するべきである．
4. **安全で安定している**
   攻撃者にシステムを踏み台にされないように攻撃を検出し続ける必要がある．

本研究では，こうした必要な事項を満たすクライアント型ハニーポットを用いて，効率のよいマルウェアの収集を目指す．

2.3.2 製品としてのハニーポット

ハニーポットには，製品として提供されているものもある．株式会社フォーティーンフォティ技術研究所 [14] は Origma+[15] という Web 感染型マルウェア検知・アラートシステムを提供している．この製品は，図 2.3 にて示すように，ユーザが指定した特定の Web サイトを定期的に巡回し，Web 感染型のマルウェアを効率的に発見・通知する人柱型のハニーポットである．また，この製品は次に述べる 4 つの状況にて利用されることが想定されている．1 つ目は，企業公開 Web サイト管理者に向けてである．2 つ目は，Web ホスティング事業者に向けてである．3 つ目は，企業社内の IT 管理者に向けてである．そして 4 つ目は，運用監視サービス事業者に向けてである．これら 4 つはすべて自社 Web サイトの改ざんの早期発見や，従業員のマルウェア感染を防ぐもので，主に企業を対象とした製品である．他の企業が提供する製品も企業向けのものであることがほとんどであり，一般の研究者が契約・導入することはコストの面からもほぼ不可能である．

2.3.3 低対話型ハニーポットを使用した事前調査

本研究を行う事前調査として，2011 年 6 月 4 日から 2012 年 11 月 12 日までの約 1 年半の期間にて，低対話型ハニーポットである Nepenthes を使用しマルウェアの収集を行った．NTT 東日本 [16]，NTTcommunications[17] の提供する OCN 光 with フレッツ ファミリータイプを利用し，一般回線上に Nepenthes を設置し調査を行った．なお，一般ユーザが被害にあう状況を想定していたため，IP アドレスは 1 つのみ割り当てている．
収集結果としては，表 2.1 の通りである．約 1 年半の間調査を行ったが，わずか 61 の検体

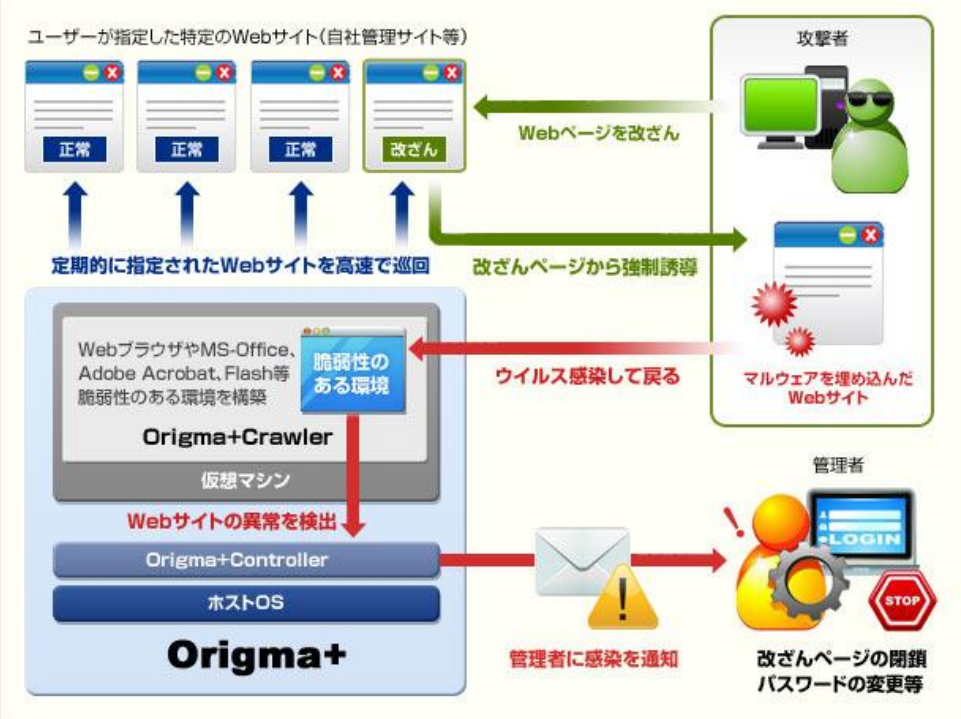


図 2.3: 株式会社フォーティーンフォティ技術研究所による Web 感染型マルウェアを Origma+が検知・アラートする仕組み

しか収集することができず，能動型攻撃の情報を収集する低対話型ハニーポットでは効率よくマルウェア検体を収集することができないということが分かる．なお，この 61 検体を VirusTotal[18] を用いて調べたところ，攻撃の足がかりとするためのバックドア型マルウェアが大半を占めていた．このことから，低対話型ハニーポットは多様な検体を収集するのに用途に適していないことが分かる．さらに，低対話型ハニーポットでは感染経路の多くを占める Drive-by download attack 型のマルウェアを収集することはできない．
こうした結果を受け，本研究では効率よくマルウェアを収集するため，受動型攻撃の情報を収集する高対話型のハニーポットを設計・実装する．

表 2.1: Nepenthes 使用による収集結果

データ収集期間	収集攻撃数	収集検体種類
2011.6.4∼2012.11.12	990 回	61 種類

2.4 本論文での着眼点

上記のとおり，マルウェアの増加および多様化が脅威であるという現状がある．そうした脅威を排除するためにマルウェアの検知・解析が欠かせない．マルウェアの検知・解析を効率化する研究を支えるためにマルウェアの検体の収集は必要不可欠である．本論文ではマルウェアの増加のスピードに対応するために，高対話型ハニーポットを利用した効率的なマルウェア収集環境を提案する．

2.5 まとめ

本章では，マルウェアによる脅威とその種類を示し，現在行われている対策およびその対策を行うために必要なマルウェアの解析・収集方法について述べた．その中でも，すべての根幹となる収集の面に注目する．マルウェアの対策について研究する研究者に情報提供を行うことができるように，これまでに存在するものよりも効率的なマルウェア収集環境の構築を目指す．

第3章　関連研究

本章では，マルウェアの効率的な収集に関する既存研究について述べる．また，悪意のある Web サイトの発見方法について言及する既存研究についても述べる．

3.1　マルウェアの収集に関する研究

ここでは，マルウェアを効率的に収集する方法について述べている既存研究を分析，記述する．

3.1.1　自律型クライアントハニーポットの提案

星澤らの研究 [19] では，インターネット上の Web サイトの中からマルウェアが存在する可能性のあるサイトを見つけて巡回（クローリング）することで，マルウェアの収集を行っている．その際に氏らの論文では，悪意のある Web サイトの URL の特徴を分析し，アクセスする URL の優先度を決めている．そして，優先度の高い URL からアクセスすることで効率よくマルウェアを収集する手法を提案している．氏らの論文では以下の項目を優先的にアクセスするための条件としている．

- **拡張子が asp，aspx，php，cgi のもの**
  これらの拡張子を持つ Web サイトは，動的にページを生成していたりデータベースと連携していたりする可能性が高い．そのため，SQL インジェクション攻撃にあう可能性が高いとし，優先的にアクセスしている．
- **ドメインが IP アドレスのもの**
  ドメイン名に IP アドレスが使用されている URL は，不正アクセスに使用される悪質サイトである可能性が高いという調査結果に基づき，優先的にアクセスしている．
- **トップレベルドメインが特定の国コードのもの**
  「.tk」（トケラウ）「.ro」（ルーマニア）「.ru」（ロシア）などのトップレベルドメインが悪質なサイトに多くみられるという調査に基づき，これらの国コードがトップレベルドメインに含まれる URL に優先的にアクセスしている．

氏らの論文では以上の 3 つの条件のいずれかを含む URL に優先的にアクセスすることで短時間に多くの URL にアクセスすることができることが示されている．しかし，氏ら

の論文では，従来のクライアント型ハニーポットと比較してどの程度効率がよくなったのか明記されておらず，評価が不十分である．また，優先的にアクセスするための条件も 3 つのみと少なく，効率化の程度に関して疑問が残る．

3.1.2 Design and implementation of high interaction client honeypot for drive-by-download attacks

秋山らの研究 [13] では，Marionette と呼ばれるクライアントハニーポットを構築しマルウェアを収集している．氏らの論文では，別の悪意のある Web サイトにリダイレクトを行う Web サイトや，脆弱性を含む Web サイト，難読化された Javascript を含む Web サイトを悪意のある Web サイトとしている．これらの悪意のある Web サイトを探して巡回することで，マルウェアを収集し，攻撃 Web サイト間の構成を調査している．
この Marionette を用いた研究の結果については，彼らの執筆した別の論文，能動的攻撃と受動的攻撃に関する調査および考察 [20] でも述べられている．この論文によると Marionette では，マイクロソフト社 [21] の提供する 31,234 の悪性 URL リストを巡回し，受動的攻撃の調査データを収集している．2008 年の 1 月 22 日から 27 日までシステムを動かした結果，全体の 10.9 %である 3,408 の URL から攻撃を検知し，9,533 の検体を収集することができている．しかし，SHA1 のハッシュ値を元に区別した結果，検体の種類は 136 種類にとどまっている．なお秋山らは，検出精度と多様性・多様な検体の収集・効率的なパフォーマンス・安全で安定していることをクライアント型ハニーポットに必要な事項としている．

3.2 悪意のある Web サイトの発見方法に関する研究

ここではどのような方法でマルウェアの含まれる悪意ある Web サイトと一般的な Web サイトを分類するか述べた既存研究について分析し，記述する．

3.2.1 Searching structural neighborhood of malicious URLs to improve blacklisting

秋山らの研究 [4] では，悪意のある Web サイトは，同ドメイン内の別のパスにも存在する可能性があるということが述べられている．この論文では悪意のある Web サイトの URL を検索エンジンで検索し，その結果を基にクローリングを行うことで，同ドメイン内での別の悪意のある Web サイトを取得している．図 3.1 に示す 2010 年 12 月 20 日の調査結果では，12,866 の悪意のある Web サイトの同ドメイン内の別のパスから 54,677 の Web サイトを取得し，その中から 278 の悪意のある Web サイトを発見することができている．

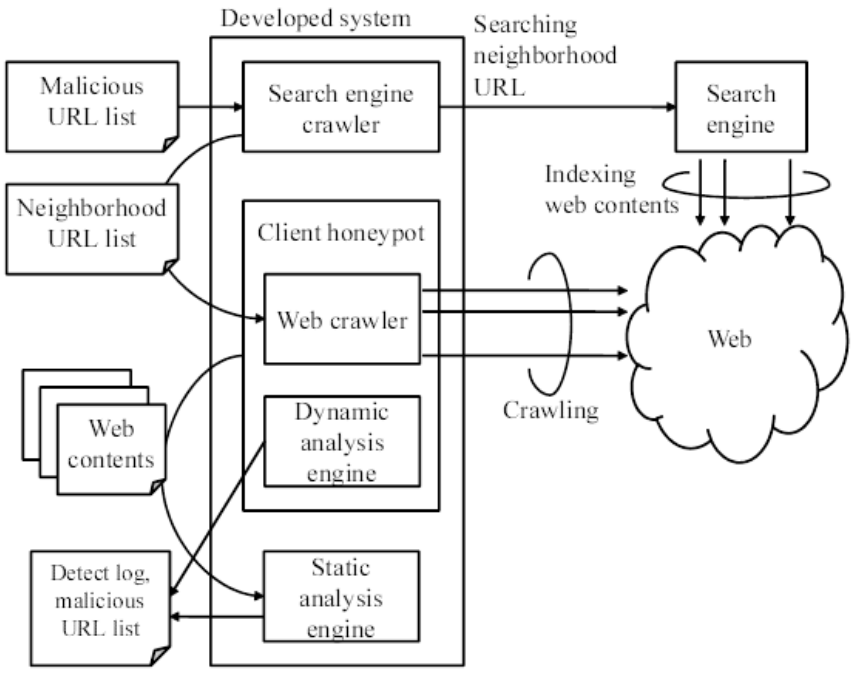


図 3.1: Searching structural neighborhood of malicious URLs to improve blacklisting, 2011, Mitsuaki Akiyama

3.2.2 Identification of Malicious Web Pages with Static Heuristics

C.Seifert らの研究 [22] では，ActiveX に含まれる脆弱性をターゲットにした exploit や exploit を呼びだす HTML のコード，リダイレクトを行う HTTP のレスポンスや難読化されている HTML のコードを含む Web サイトを，悪意のある Web サイトとしている．これらのデータを入力データとして機械学習を行うことで，悪意のある Web サイトと正常な Web サイトを分類している．C.Seifert らの論文では，図 3.2 のような決定木という手法を用いることで 61,000 個の URL を 49 分間で巡回できることが示されている．この論文では，5,678 の悪意のある Web サイトと 16,006 の正常な Web サイトを訓練データとし，機械学習を行っている．その結果，61,000 個の URL の中から 3,590 個の URL を悪意のある Web サイトと判定しているが，5.88 %のフォールスポジティブと 46.15 %のフォールスネガティブが生じている．このことから，この論文における手法では，誤検知の割合を低くすることが可能である代わりに検知漏れが多くなってしまうという結果が生じている．

3.2.3 検知を目指した不正リダイレクトの分析

寺田らの研究 [23] では，Drive-by download attack における Web ページへのアクセスの遷移に着目し，そのアクセス履歴の特徴を明らかにしている．また，機械学習の決定

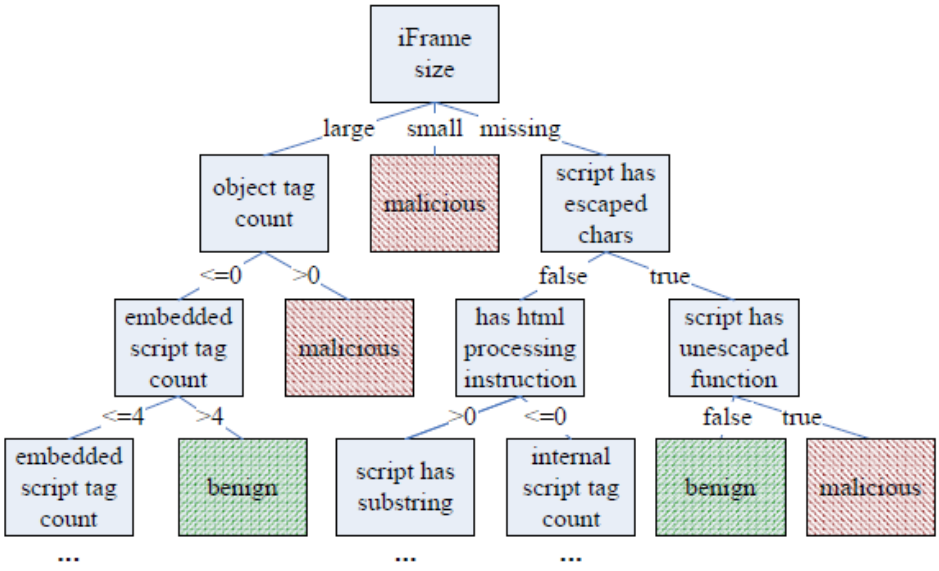


図 3.2: Identification of Malicious Web Pages with Static Heuristics, 2008, Christian Seifert

木学習手法を用いてマルウェアを配布する悪意のある Web サイトを抽出している．この論文では，攻撃通信データから HTTP 通信セッションを再構築し，マルウェアのダウンロードに相当する危険なアクセスを抽出し，その HTTP リクエストに至る遷移の特徴を明らかにしている．氏らの論文では，PDF ファイル，SWF ファイル，バイナリファイルの 3 種類のダウンロードが危険であると判断し，HTTP レスポンスの Content-Type ヘッダを参照し危険なレスポンスを特定している．リクエストの種類とこれらの関係は図 3.3 に示されている．この結果により，不明度の高いリクエストに危険なレスポンスがある可能性が高いことが示されている．

また，氏らは巡回対象の URL から別の URL までの距離や URL の出すリクエスト数，送信元送信先のパケット数，データ数などを入力データとし，機械学習を行うことで悪意のある Web サイトの URL を予測した．その結果，悪意のある Web サイトを正しく判断できた割合は 85.1%と高い値を示せたものの，FN が 12.3%，FP が 17.1%と決して低いとは言えない結果が生じている．

	総数	レスポンス		
		pdf	swf	bin
Location	199	0	0	0
		0.00%	0.00%	0.00%
Url In Data	737	13	8	40
		1.76%	1.09%	5.43%
Path In Data	1407	0	8	44
		0.00%	0.57%	3.13%
Equal Host	3296	183	88	521
		5.55%	2.67%	15.81%
other	731	79	1	8
		10.81%	0.14%	1.09%
総数	6370	275	105	613

図 3.3: リクエスト種類と危険レスポンスの関係，2010，寺田 剛陽

3.2.4 The Ghost In The Browser Analysis of Web-based Malware

N.Provos らの研究 [24] では，Google のクローラによって収集された Web ページの中から悪意のある Web ページを自動抽出し，悪意のある Web ページを作成する側の戦略や傾向について調査を行っている．この論文では，別の悪意のある Web ページに対して iframe を用いたリンクを張っていたり，難読化を施された JavaScript を含んでいる Web ページを悪意のある Web ページとし，MapReduce を用いてフィルタリングを行うことで抽出している．氏らの論文では，図 3.4 のように，多いときで 1 日で 1 万から 3 万の悪意のある Web ページの URL を探し出すことができ，1 日当たり 30 万の URL を処理することが可能になっている．
この論文では 450 万の URL の分析を行い，その内の 10 %に当たる 45 万の URL が Drive-by download attack に関わっていたとしている．

3.3 まとめ

本章では，マルウェア検体の収集に関する研究と悪意のある Web サイトの発見方法に関する研究の，2 つの種類の研究について紹介し，その手法について述べた．これらの関連研究より，マルウェア検体を収集する際に，悪意のある Web サイトの URL を起点とし Web クローリングを行うことが効率的であることが分かっている．さらに，悪意のある Web サイトの情報を基に，悪意のある Web サイトである可能性の高いサイトを探すことによって，より早く沢山のマルウェアの検体を収集できることが分かっている．しかし，これらの研究ではフォールスネガティブ，つまり検知漏れが多くなっている．そのため，

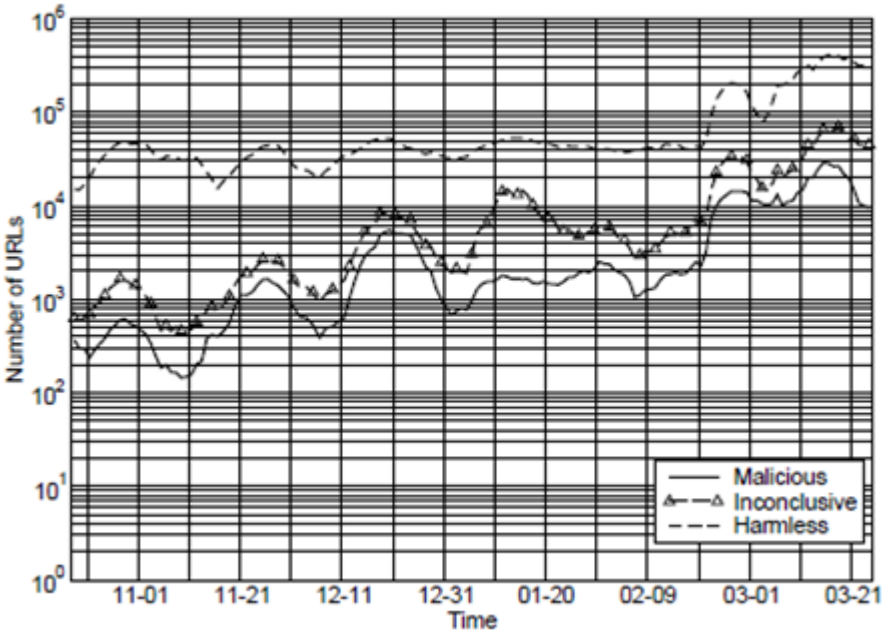


図 3.4: The Ghost In The Browser Analysis of Web-based Malware, 2007, Niels Provos

速さと検体数の多さという効率のよさを保ちつつ，より正確に機能する手法が必要であるといえる．

第4章　システム設計

2012年第1四半期の段階で，PCを攻撃するマルウェアの数は近年の中で最も多い状況にあるということを第2章で述べた．さらに，第2章で述べたように，現存するマルウェア検体の挙動を解析し対策をとることが非常に重要である．こうした対策をとるためにも，継続的なマルウェア検体の収集を迅速かつ効率的に行う必要がある．本研究では，受動的な攻撃を行う悪意のあるWebサイトに自らアクセスし，マルウェアを収集する能動型ハニーポットを提案し，構築する．本章では，まず前提として，受動的な攻撃を行う悪意のあるWebサイトを見つける方法について述べる．その次に，発見した悪意のあるWebサイトを種（Seed）としてWebクローリングを行い，悪意のあるWebサイトである可能性の高いWebサイトを巡回する方法について述べる．最後に，収集したWebサイトの中から実際に悪意のあるWebサイトである可能性が高いWebサイトを見つけて検体を収集する方法について述べる．

4.1　前提

能動的にマルウェアを収集するために，まず，受動的な攻撃を行う悪意のあるWebサイトを特定することが必要である．本研究では以下の方法でそうしたWebサイト,「種（Seed）となるWebサイト」を収集している．

- **Malware Domain List**
  Malware Domain List[25]とは，図4.1のように悪意のあるWebサイトの情報を集めて掲載する，非商用のプロジェクトである．このWebサイトでは，様々な国からの通報をまとめて掲載しているため更新が不定期であり，1日の間に0件から10件程度の悪意のあるWebサイトの情報が更新されている．
  本研究では，1時間に一度Malware Domain Listの更新を確認し，更新があった際にそのURLを取得し，Seedとなる URLとしている．

- **Malware Blask List**
  Malware Black List[26]も，Malware Domain Listと同様に，図4.2のように悪意のあるWebサイトの情報を集めて掲載している．このサイトは多い日では1日に100件を超える悪意のあるWebサイトの情報を更新している．
  本研究では，1時間に一度Malware Black Listの更新を確認し，更新があった際にそのURLを取得し，Seedとなる URLとしている．

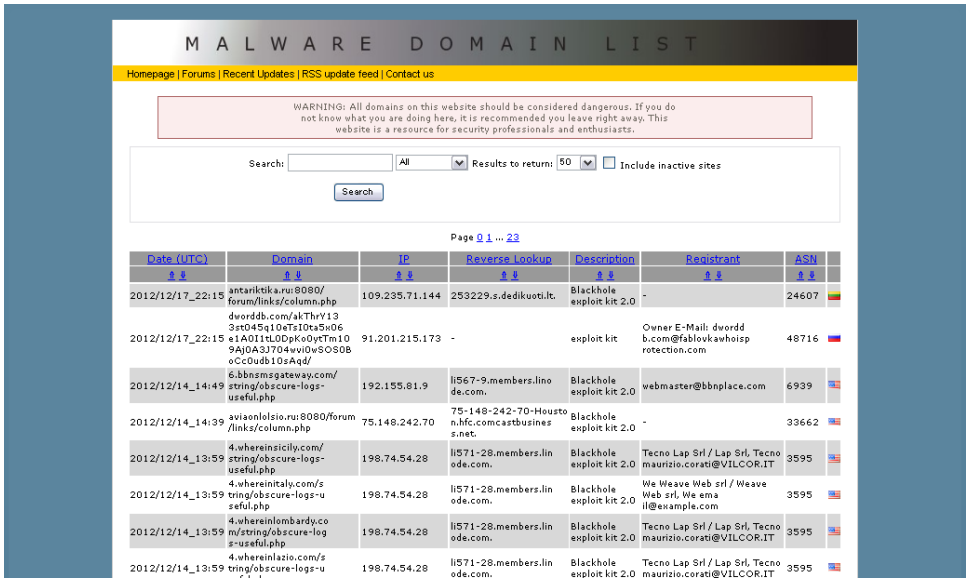


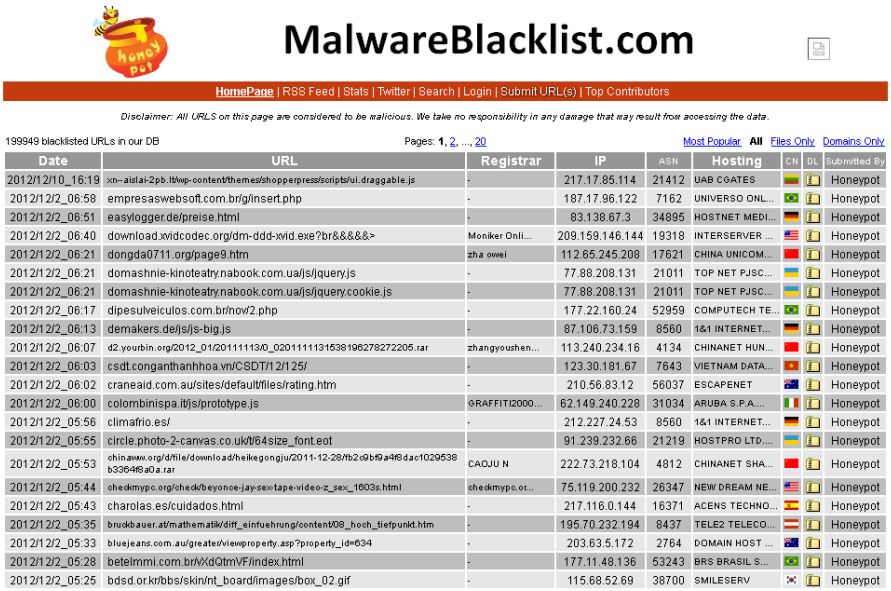
図 4.1: Malware Domain List(http://www.malwaredomainlist.com/)

4.2 Web クローリング

第 3 章で述べた秋山らの研究 [4] によると，マルウェアを配布する悪意のある Web サイトは同ドメイン内の異なるパスに存在する可能性が高いということが分かっている．そのため，本研究では前項で述べた Seed となる URL をもとにして Web クローリングを行うことで，効率よく悪意のある Web サイトを収集することを目的とする．Web クローリングを行う際にはまず，Web サイトの情報を取得する．そして，その Web サイトに他のサイトへのリンクがあった場合，そのリンク先の URL を取得しさらにその Web サイトの情報を取得する．この手順を繰り返すことで効率よく Web サイトの URL を収集することができる．しかし，Web クローリングを行って集めた URL は悪意のある Web サイトのものばかりではない．したがって，短時間でより多くの悪意のある Web サイトにアクセスするために，悪意のある Web サイトである可能性が高いサイトから優先的にアクセスすることが必要である．

4.3 悪意のある Web サイトの判定方法

前項で述べたように，Web クローリングを行い収集した Web サイトの URL は必ずしもすべてが悪意のある Web サイトのものではない．そのため，Web クローリングを行い収集した URL すべてにアクセスすることは，効率的なマルウェア検体の収集にはつなが



MalwareBlacklist.com

HomePage | RSS Feed | Stats | Twitter | Search | Login | Submit URL(s) | Top Contributors

Disclaimer: All URLs on this page are considered to be malicious. We take no responsibility in any damage that may result from accessing the data.

199949 blacklisted URLs in our DB　　Pages: **1**, 2, ..., 20　　Most Popular　**All**　Files Only　Domains Only

Date	URL	Registrar	IP	ASN	Hosting	CN	DL	Submitted By
2012/12/10_16:19	xn--aislai-2pb.it/wp-content/themes/shopperpress/scripts/ui.draggable.js	-	217.17.85.114	21412	UAB CGATES			Honeypot
2012/12/2_06:58	empresaswebsoft.com.br/g/insert.php	-	187.17.96.122	7162	UNIVERSO ONL...			Honeypot
2012/12/2_06:51	easylogger.de/preise.html	-	83.138.67.3	34895	HOSTNET MEDI...			Honeypot
2012/12/2_06:40	download.xvidcodec.org/dm-ddd-xvid.exe?br&&&&&>	Moniker Onli...	209.159.146.144	19318	INTERSERVER ...			Honeypot
2012/12/2_06:21	dongda0711.org/page9.htm	zha owei	112.65.245.208	17621	CHINA UNICOM...			Honeypot
2012/12/2_06:21	domashnie-kinoteatry.nabook.com.ua/js/jquery.js	-	77.88.208.131	21011	TOP NET PJSC...			Honeypot
2012/12/2_06:21	domashnie-kinoteatry.nabook.com.ua/js/jquery.cookie.js	-	77.88.208.131	21011	TOP NET PJSC...			Honeypot
2012/12/2_06:17	dipesulveiculos.com.br/now/2.php	-	177.22.160.24	52959	COMPUTECH TE...			Honeypot
2012/12/2_06:13	demakers.de/js/js-big.js	-	87.106.73.159	8560	1&1 INTERNET...			Honeypot
2012/12/2_06:07	d2.yourbin.org/2012_01/20111113/0_020111113163819627822205.rar	zhangyoushen...	113.240.234.16	4134	CHINANET HUN...			Honeypot
2012/12/2_06:03	csdt.conganthanhhoa.vn/CSDT/12/125/	-	123.30.181.67	7643	VIETNAM DATA...			Honeypot
2012/12/2_06:02	craneaid.com.au/sites/default/files/rating.htm	-	210.56.83.12	56037	ESCAPENET			Honeypot
2012/12/2_06:00	colombinispa.it/js/prototype.js	GRAFFITI2000...	62.149.240.228	31034	ARUBA S.P.A....			Honeypot
2012/12/2_05:56	climafrio.es/	-	212.227.24.53	8560	1&1 INTERNET...			Honeypot
2012/12/2_05:55	circle.photo-2-canvas.co.uk/t/64size_font.eot	-	91.239.232.66	21219	HOSTPRO LTD:...			Honeypot
2012/12/2_05:53	chinawww.org/d/file/download/heikegongju/2011-12-28/fb2c0bf0a4f6dac1029538b33648a0a.rar	CAOJU N	222.73.218.104	4812	CHINANET SHA...			Honeypot
2012/12/2_05:44	checkmypc.org/check/beyonce-jay-sex-tape-video-z_sex_1603s.html	checkmypc.or...	75.119.200.232	26347	NEW DREAM NE...			Honeypot
2012/12/2_05:43	charolas.es/cuidados.html	-	217.116.0.144	16371	ACENS TECHNO...			Honeypot
2012/12/2_05:35	bruckbauer.at/mathematik/diff_einfuehrung/content/08_hoch_tiefpunkt.htm	-	195.70.232.194	8437	TELE2 TELECO...			Honeypot
2012/12/2_05:33	bluejeans.com.au/greater/viewproperty.asp?property_id=634	-	203.63.5.172	2764	DOMAIN HOST ...			Honeypot
2012/12/2_05:28	betelmmi.com.br/vXdQtmVF/index.html	-	177.11.48.136	53243	BRS BRASIL S...			Honeypot
2012/12/2_05:25	bdsd.or.kr/bbs/skin/nt_board/images/box_02.gif	-	115.68.52.69	38700	SMILESERV			Honeypot

図 4.2: Malware Black List(http://www.malwareblacklist.com)

らない．本研究では，集めた悪意のある Web サイトの可能性がある Web サイトの URL を，第 4.3.1 項にて後述する独自の判定基準を用いることで悪意のある Web サイトのものかどうかを判定する．判定を行うに当たり，決定木学習という機械学習手法を用いる．その結果悪意のある Web サイトのものである可能性が高いと判定された，Web サイトの URL から優先的にアクセスを行うことで効率的なマルウェア検体の収集を実現する．

4.3.1　悪意のある Web サイトの判定基準

ここでは，先述した本研究独自の判定基準について述べる．以下に悪意のある Web サイトと見なすための判定基準を示す．

1. **トップレベルドメインが特定のもの**
   Malware Domain List 及び Malware Black List に掲載されている情報から，悪意のある Web サイトに多くみられるトップレベルドメインの特徴を推測し，そのトップレベルドメインを含む URL の Web サイトを悪意のあるものとみなす．Malware Domain List に掲載されている悪意のある Web サイトの URL に多く含まれるトップレベルドメインを表 4.1 に記す．また，Malware Black List に掲載されている悪意のある Web サイトの URL に多く含まれるトップレベルドメインを表 4.2 に記す．

   表 4.1 及び表 4.2 をみると，「.com」「.net」「.biz」といった，安価で誰もが取得しやすいトップレベルドメインが多く含まれていることがわかる．また，「.cn」「.br」「.ru」

表 4.1: Malware Domain List にて多くみられたドメイン (上位 10 個)

	ドメイン	数
1	com	30889
2	ru	6362
3	net	5769
4	cn	5418
5	info	5346
6	in	4346
7	cc	4118
8	org	2511
9	biz	1269
10	br	978
	総数	77592

表 4.2: Malware Black List にて多くみられたドメイン (上位 10 個)

	ドメイン	数
1	org	271
2	de	191
3	pl	161
4	cn	126
5	uk	122
6	in	81
7	br	73
8	info	72
9	it	68
10	kr	65
	総数	2000

「.in」といった BRICs の国々のような経済発展が著しい国のトップレベルドメインが多く含まれていることがわかる.

2. **ドメイン名が設定されていないもの**
悪意のある Web サイトは，特定されて，ブラックリストに掲載されることを防ぐため，生存している時間が短い場合が多く見られる．そのため，すぐにサイトの閉鎖や URL の変更ができるように，ドメインを取得せず IP アドレスのみを含む場合が

多いと考えられる．

3. **whois に含まれる国情報が特定の国のもの**
Malware Domain List 及び Malware Black List に掲載されているドメインに対して whois コマンドを実施する．悪意のある Web サイトに多くみられる国情報の特徴を抽出し，それを含む Web サイトの URL を悪意のあるものとみなす．
Malware Domain List に掲載されている悪意のある Web サイトに多く見られる国別コードトップレベルドメイン（以下，ccTLD とする）を図 4.3 に記す．また，Malware Black List に掲載されている悪意のある Web サイトに多く見られる ccTLD を図 4.4 に記す．

表 4.3: Malware Domain List にて多くみられた ccTLD(上位 10 個)

	ccTLD	数
1	CN	6400
2	RU	1296
3	NL	1027
4	BR	851
5	DE	772
6	UA	760
7	EU	598
8	TR	579
9	KR	411
10	LV	550
	総数	19674

4. **URL 中に blog,page,wiki という文字を含まないもの**
Web クローリングを行う際に blog や page，wiki といった単語を含む場合，同ドメイン内のすべての Web サイトにアクセスしてしまうと，情報量が膨大になりすぎてしまう．そのため，効率化が図れなくなるということが問題となる．
そのため，URL 中に blog，page，wiki という文字が含まれている場合，そのサイトは Web クローリングの対象には含まないものとする．

5. **別のサイトへリダイレクトを行うもの**
Malware Domain List 及び Malware Black List に掲載される情報より，HTTP ヘッダ情報に location が含まれるものが多くあることが分かった．location は別のサイ

表 4.4: Malware Black List にて多くみられた ccTLD(上位 10 個)

	ccTLD	数
1	DE	183
2	CN	146
3	PL	83
4	NL	50
5	FR	45
6	BR	35
7	IT	32
8	GB	26
9	TR	21
10	ES	20
	総数	903

トへのリダイレクトが行われていることを示すものであり，悪意のある Web サイトに多くみられる特徴であると推測する．そのため，HTTP ヘッダ情報に location 情報を含む Web サイトを悪意のある Web サイトであるとする．

6. **HTTP ヘッダ情報に特定の情報が含まれるもの**
   Malware Domain List 及び Malware Black List に掲載される情報より，HTTP ヘッダに含まれるサーバの情報に x-server の使用が記載されているものが多くみられた．そのため，HTTP ヘッダでの x-server の使用の記載は悪意のある Web サイトに多くみられる特徴であると推測する．

7. **IP アドレスが特定のもの**
   悪意のある Web サイトは，同一の攻撃者が何度も形を変えて作成している可能性があるため，同一の IP アドレスが使用されることがあると推測される．Malware Domain List 及び Malware Black List に記載されている悪意のある Web サイトの IP アドレスと一致する IP アドレスをもつ Web サイトを悪意のある Web サイトのものであるとする．本研究での調査の結果，第 2 オクテットまでは同じであるが第 3 オクテット以降が異なる IP アドレスが多く見られたため，第 1 オクテット及び第 2 オクテットの情報のみを参考としている．

8. **特定のレジストラに登録しているもの**
   悪意のある Web サイトの URL に対して whois コマンドを実施し，レジストラ情報

を取得することで，悪意のある Web サイトが多く登録するレジストラを特定する．Malware Domain List に掲載されている悪意のある Web サイトに多く見られるレジストラ情報を表 4.5 に記す．また，Malware Black List に掲載されている悪意のある Web サイトに多く見られるレジストラ情報を表 4.6 に記す．

表 4.5: Malware Domain List にて多くみられたレジストラ情報 (上位 10 個)

	レジストラ	数
1	SMA4	1249
2	THEPL	1004
3	TECHN33	1004
4	ABUSE271	1004
5	IPADM258	701
6	NOC124	686
7	NETWO1546	557
8	TPCM	546
9	LNO21	494
10	ABUSE1025	469
	総数	9854

表 4.6: Malware Black List にて多くみられたレジストラ情報 (上位 10 個)

	レジストラ	数
1	NOC124	58
2	ZD69	54
3	DAT5	39
4	THEPL	38
5	TECHN33	38
6	ABUSE271	38
7	ABUSE51	29
8	MCRAE6	24
9	NDN	24
10	HNI1	22
	総数	891

以上で述べた判定基準のうち，いくつかに該当する情報を持つ Web サイトを悪意のある Web サイトのものであるとする．その判定基準の重要度は，判定基準を学習データと

した機械学習を行うことで特定する．機械学習の手法は決定木学習を用いる．決定木学習を用いた判定については第 4.3.2 項にて述べる．

4.3.2 決定木学習を用いた悪意のある Web サイトの判定手法

悪意のある Web サイトの特徴を用いて判定を行うに当たり，どのようにして特徴情報を組み合わせ，判定を行うかとが重要である．そのため，本研究では第 4.3.1 項にて述べた判定基準を学習データとした決定木学習を行った．決定木学習を用いた悪意のある Web サイトの判定について以下の 3 つの点を述べる．1 つ目は決定木学習とは何かについて述べる．2 つ目は決定木学習の利点について述べる．3 つ目は本研究における決定木学習による判定について述べる．

決定木学習

決定木学習とは，機械学習の分野における予測モデルである．また，ある事項に対する観察結果から，その事項の目標値に関する結論を導くものである．決定木は図 4.3 に示す木構造をとり，ノード（節点，頂点）とノード間を結ぶエッジ（枝，辺）あるいはリンクにて表される．ノードには何らかのデータ（値，条件）が付属している．どのような入力データも，木構造のトップである根ノードから始まり，各ノードの判定基準に従いながら特定の葉ノードに落ちる経路をたどる．葉ノードとは，根ノードからの経路により表わされる変数値に対する予測値を表している．

決定木学習の利点

次に，悪意のある Web サイトを判定するにあたり，決定木学習を利用した理由について述べる．決定木学習を使用する利点として，3 つの点があげられる．1 つ目はデータの処理が必要ない点である．決定木学習では，非計量的なデータを扱うことができる．そのため，本研究にて使用する，判定基準に含まれる国情報をビットベクトル化せず扱うことができる．非計量的なデータをビットベクトル化した場合，次元が大きくなり，過学習が起きる可能性がある．2 つ目は超平面では分割が困難である状況でも適用が可能な点である．超平面では分割が困難である状況を図 4.4 に示す．3 つ目は分析結果の評価や解釈が容易な点である．分析結果のモデルを作成することができるため，どのデータを基準に判定を行っているのかを特定することが可能である．以上の 3 つの利点により，本研究では決定木学習を用いる．

決定木学習を用いた判定

最後に，本研究における決定木学習による判定について述べる．判定を行う前に，既知の悪意のある Web サイトおよび正常な Web サイトの URL を訓練データとして学習させ，

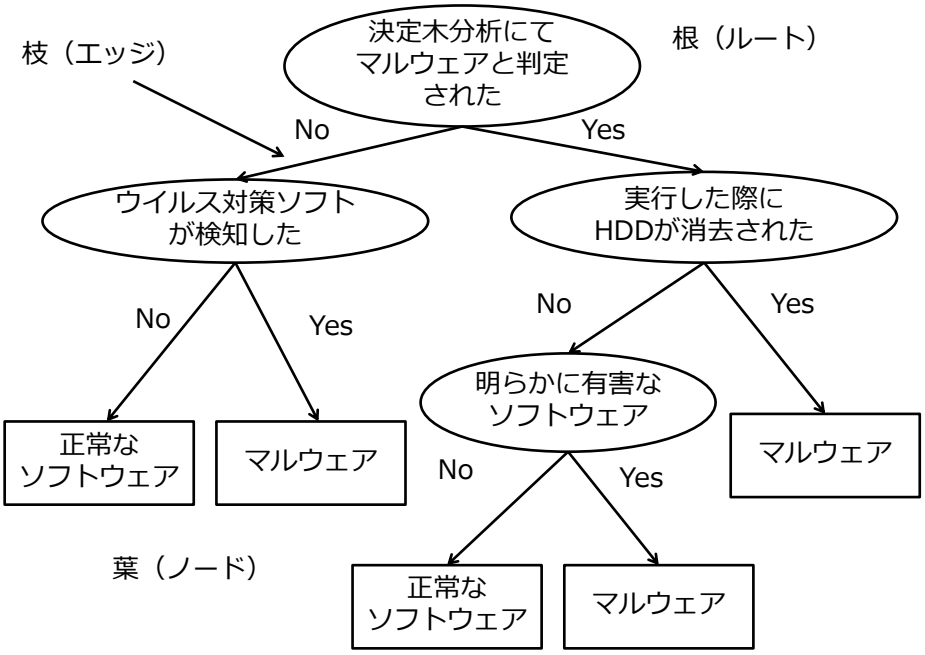


図 4.3: 木構造:マルウェア判別モデル

判定のための準備を行う．まず，訓練データとなる URL 情報は第 4.3.1 項にて述べた項目に該当するかどうかということをそれぞれ確認する．そして，訓練データから 17 次元の特徴ベクトルを構成し，学習させる．17 次元の特徴ベクトルについては以下の表 4.7 にて示す．これにより，入力データを使用した判定が可能となる．入力データのデータ構造を以下の図 4.5 に示す．

決定木学習を行うことにより，Web クローリングにて収集したすべての Web サイトを巡回することなく Web サイトの URL を収集することができる．そのため，SeedURL が増加するにつれ，判定を行わない際と比較して Web クローリング速度の上昇が期待される．

4.4 まとめ

本章では，マルウェア検体を収集するうえで前提となる SeedURL の種類について述べ，受動型攻撃の情報を収集するために必要な Web クローリングの方法について述べた．また，効率的なマルウェア検体の収集を可能にするための手法について述べた．マルウェアを配布する悪意のある Web サイトに多くみられる特徴を推測し，Web クローリングを行い取得した Web サイトの URL が悪意のある Web サイトのものなのかそうでないのかを判定する．そして，判定した結果，悪意のある Web サイトである可能性が高いとされた Web サイトから優先的にクローリングを行う．これにより，短時間でより多くの Web サイトを効率的にクローリングすることを目指す．第 5 章では，本章で紹介した手法を用い

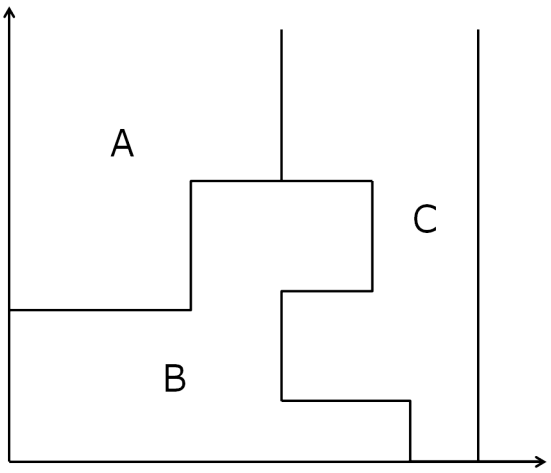


図 4.4: 超平面では分割が困難である状況

表 4.7: 特徴ベクトルの成分

成分名	内容	表示形式
IP-1	IP アドレスの第 1 オクテット	1~255 の任意の数字
IP-2	IP アドレスの第 2 オクテット	1~255 の任意の数字
Country	国情報	2 文字国コード
Domain	URL に含まれるトップレベルドメイン	TLD
location	リダイレクトの有無	0(含まない) もしくは 1(含む)
x-server	server 情報	0(含まない) もしくは 1(含む)
IPAddress	URL 中の IP アドレスの有無	0 もしくは 1
SMA-ARIN	レジストラ情報	0 もしくは 1
ABUSE271-ARIN	レジストラ情報	0 もしくは 1
IPADM258-ARIN	レジストラ情報	0 もしくは 1
NETWO1546-ARIN	レジストラ情報	0 もしくは 1
LNO21-ARIN	レジストラ情報	0 もしくは 1
ABUSE1025-ARIN	レジストラ情報	0 もしくは 1
NOC2426-ARIN	レジストラ情報	0 もしくは 1
NOC	レジストラ情報	0 もしくは 1
HNI1-ARIN	レジストラ情報	0 もしくは 1
ENGIN7-ARIN	レジストラ情報	0 もしくは 1

て実装した，マルウェアを効率的に収集するシステムについて述べる．

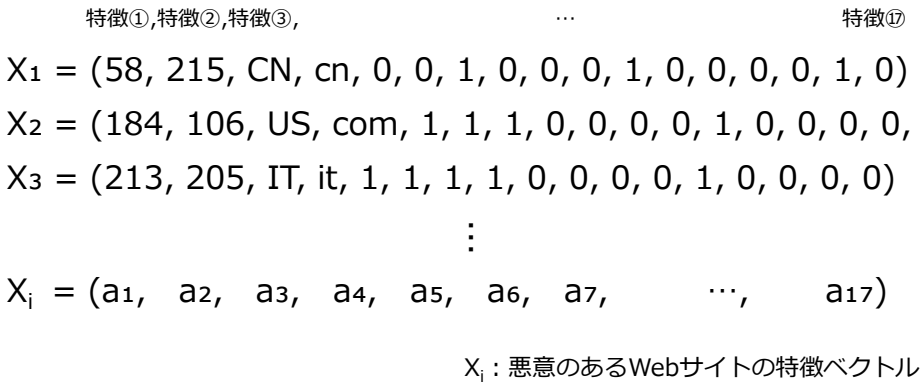


図 4.5: 入力データのデータ構造

第5章　実装

第4章で述べた手法を用いて，悪意のあるWebサイトのURLを収集し実際にアクセスすることで，受動型攻撃を行うWebサイトにて配布されるマルウェア検体を自動で収集するシステムを実装した．本章では，今回利用した実装環境を示し，そのシステム構成の詳細について述べる．

5.1　実装環境

本システムの実装環境を表5.1に示す．

表 5.1: 実装環境

要素	属性	利用環境
OS	実機	Ubuntu 11.04(32-bit)
	VirtualBox	Windows XP SP なし (32-bit)
言語	実機	Ruby1.8.7
	VirtualBox	VBScript5.6
ライブラリ	HTML 解析	nokogiri-1.5.0
フレームワーク	Web クローラ	anemone-0.6.1

5.2　実装したシステムの構成

ここでは，実装したシステムのシステム構成について述べる．本システムはSeedURL収集部，Webクローリング部，判定部，収集部の4つの部分からなる．本システムの設計を図5.1に，概要を図5.2に示す．また，以下に各部の詳細について述べる．

5.2.1　SeedURL収集部分

SeedURLの収集部分では，Malware Domain List及びMalware Black Listの2つの悪意のあるWebサイトの情報を集めたWebサイトに対してスクレイピングを行い，マル

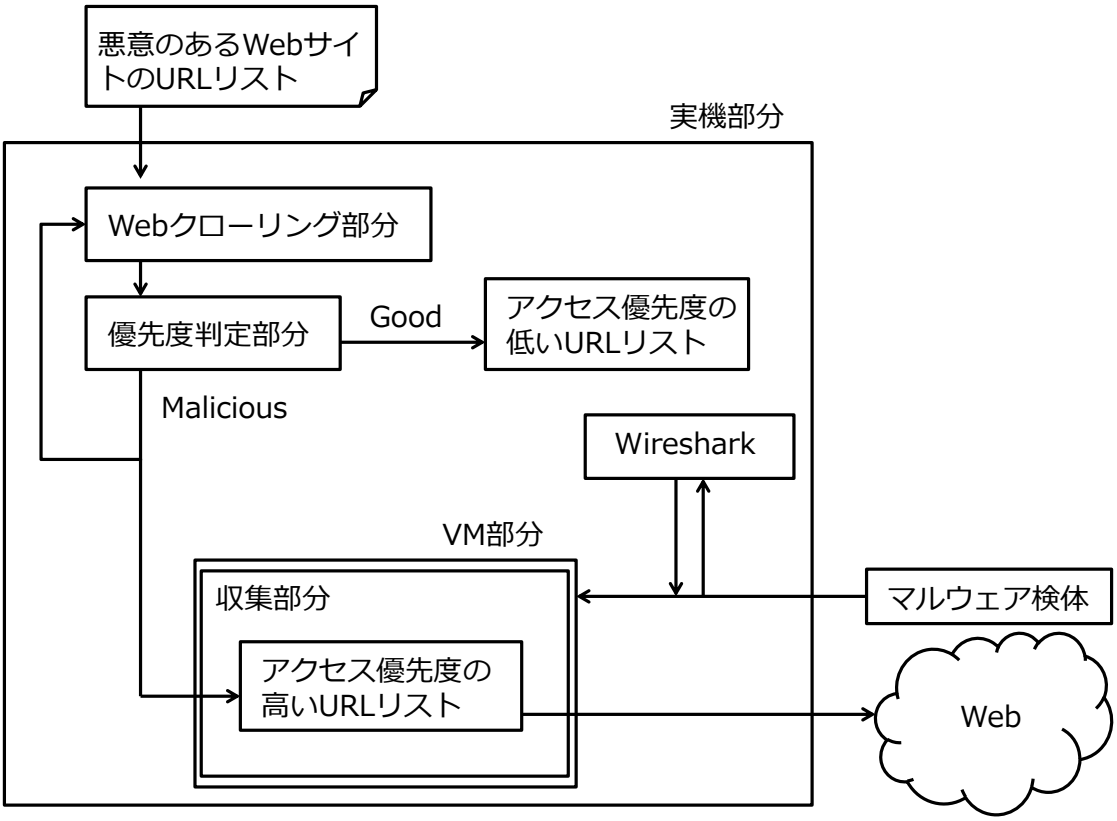


図 5.1: システム設計

ウェアを配布する悪意のある Web サイトの URL を収集する．それぞれの Web サイトから収集することのできた SeedURL の個数は表 5.2 にて示す.

表 5.2: 1 週間に収集することが可能な SeedURL 数

対象サイト	収集期間	URL
Malware Domain List	2012.12.3∼2012.12.9	69 個
Malware Black List	2012.11.25∼2012.12.2	1,457 個

5.2.2 Web クローリング部分

Web クローリングを行う部分では，先述した SeedURL の Web サイトからリンクが張られているサイトを巡回し，Web サイトの URL を収集する．Ruby のフレームワークである Anemone を使用することで Web サイトの URL を収集する．Anemone を使用することで，SeedURL の Web サイトにアクセスした際に HTML を解析し，他の Web サイトに

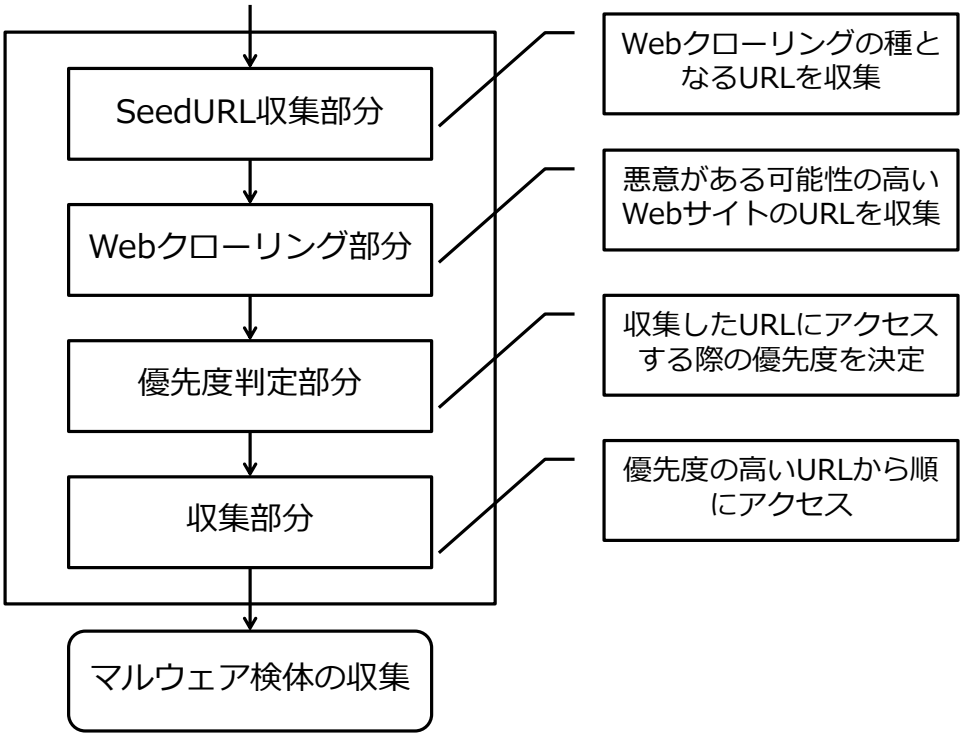


図 5.2: システム概要

リンクを張っている場合にそのリンク先の URL を収集することができる．この Web クローリングを行うことでどの程度の数の Web サイトの URL を収集できたかを以下の表 5.3 にて示す.

表 5.3: Web クローリングを行うことで収集できた URL 数

SeedURL 記載サイト	期間	SeedURL 数	Web クローリング後 URL 数
Malware Domain List	2012.7.5.∼2012.12.16	938 個	14,056 個
Malware Black List	2012.11.22∼2012.12.2	2,000 個	48,387 個

マルウェアを配布する悪意のある Web サイトは短期間で姿を消すものや，脆弱性が修正されるものが大半であり，タイミングによっては Web クローリングを行う際にもアクセスできない Web サイトも存在する．そのため，本論文ではその可能性をできるだけ減らし，有効な SeedURL を取得するために可能な限りサイトが更新された際にスクレイピ

ングを行う．さらに，この表 5.3 からわかるように，Seed となる URL の数が多ければ多いほど Web クローリングにて収集できる Web サイトの URL 数は上昇し，より悪意のある Web サイトの URL を収集できる可能性が高くなる．
しかし，Web クローリングを行うことで収集した URL はすべてが悪意のある Web サイトのものであるというわけではない．そのため，本論文にて提案する判定基準にて Web サイトの URL が悪意のあるものなのかどうか判定を行うことで悪意のある Web サイトへのみのアクセスを可能にする．判定部分については次の第 5.2.3 項にて述べる．

5.2.3 優先度判定部分

第 5.2.2 項で述べた部分にて収集した Web サイトの URL は，そのすべてが悪意のある Web サイトのものであるというわけではない．よって，収集したすべての URL にアクセスすることは，効率的ではないと判断する．そのため，Web クローリングを行い収集した URL を，悪意のある Web サイトのものかどうか第 4.3.1 項にて述べた方法にて判定する．これらの 8 種類の判定基準をもとに機械学習手法の決定木を用いることで，判定を行う．8 種類の判定基準から作成した 17 個の特徴ベクトルは先述した表 4.7 にて示す．本論文では，Malware Domain List 及び Malware Black List 掲載の 4,334 個の悪意のある Web サイトの URL を学習データとする．また，Alexa 及び Google ad planner 掲載の 1,216 個の正常な Web サイトの URL を学習データとする．判定部分の概要は図 5.3 にて示す．

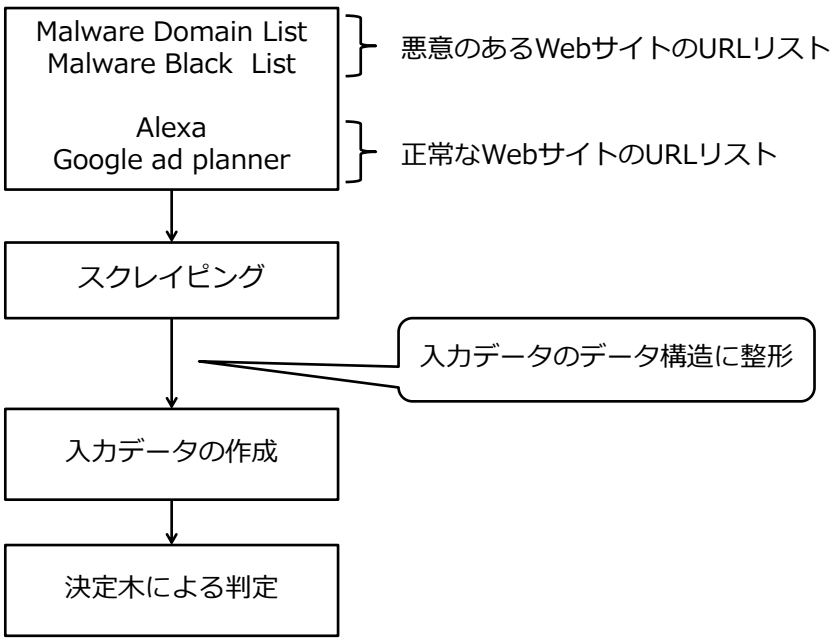


図 5.3: 判定部分概要

決定木にて悪意のある Web サイトであると判定された Web サイトから優先的にアクセスすることで効率化を図る．

5.2.4 収集部分

収集の部分は，実際にマルウェアを配布する Web サイトにアクセスする必要がある．そのため，マルウェアに感染してしまうリスクを考え，VirtualBox を用いた仮想マシン上で実行する．収集部分の概要は図 5.4 にて示す．仮想マシンの OS は脆弱性が多くみられる WindowsXP の SP なしを使用している．ブラウザは，脆弱性が多くみられる Internet-Explorer6.0 を使用している．また，仮想マシンでマルウェアを配布する悪意のある Web サイトにアクセスする際に，実機の方で Wireshark を起動しパケットキャプチャを行う．悪意のある Web サイトにアクセスする際の通信を監視することで，マルウェア検体だけでなく通信の情報も取得する．

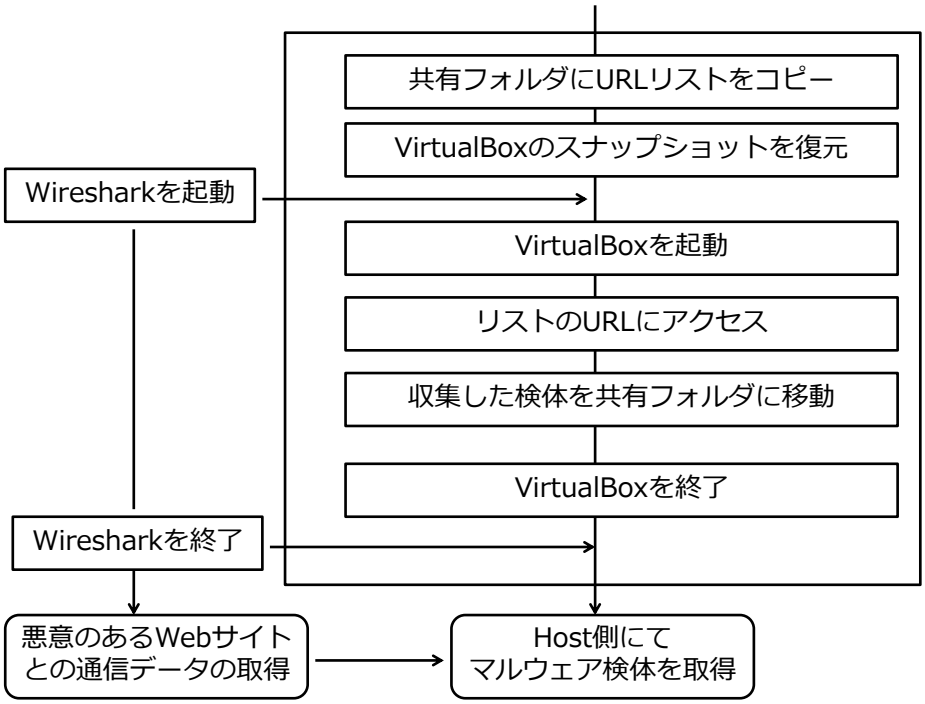


図 5.4: 収集部分概要

5.3 まとめ

本章では，第 4 章で述べた手法に基づいて実装したシステムの構成について述べた．また，システムを構成する SeedURL の収集部分，Web クローリング部分，優先度の判定部分，収集部分の 4 つの部分について詳細を明らかにした．SeedURL の収集部分では，

Malware Domain List 及び Malware Black List に対してスクレイピングを行い，悪意のある Web サイトの URL を収集する．Web クローリング部分では，Ruby のフレームワークである Anemone を使用し，SeedURL の Web サイトからリンクが張られているサイトを巡回する．優先度の判定部分では，抽出した悪意のある Web サイトの特徴を学習データとし，機械学習手法の決定木を用いて判定を行う．収集部分では，Virtual Box を用いた仮想マシン上で，優先度の判定部分にて悪意のある Web サイトであると判定された Web サイトにアクセスしマルウェア検体を収集する．第 6 章では提案する手法に基づいて設計，実装したシステムを用いて行った 3 種類の実験とその結果について述べる．

第6章　実験と結果

本章では，提案する手法に基づき設計・実装したシステムを用いて行った3種類の実験とその結果について述べる．1つ目は，本システムにおけるWebクローリングの有効性の検証実験である．Malware Domain List及びMalware Black Listに掲載されているURLをSeedとして，Webクローリングを行い，収集したURLにはどのくらいの割合で悪意のあるWebサイトのものが含まれているのか検証した．2つ目は，独自の判断基準に基づいた悪意のあるWebサイトの判定実験である．この実験では，本システムを用いることで判定を行い，悪意のあるWebサイトと正常なWebサイトを正しく判定できるか検証した．3つ目は，本システムを使用したマルウェア検体の収集実験である．この実験では，Malware Domain List及びMalware Black Listに掲載されているURLをSeedとして，実装したシステムを実際に動かすことでどの程度のマルウェア検体を収集することができるのか検証した．

6.1　Webクローリングの有効性検証実験

本システムを用いて実際にWebクローリングを行い，収集したURLに含まれる悪意のあるWebサイトのURLの割合を検証する実験を行った．以下にこの実験の概要と実験環境を提示し，その結果を記述する．

6.1.1　実験概要

マルウェア検体を効率的に収集するためには，マルウェアを配布する悪意のあるWebサイトのURLを効率的に収集する必要がある．さらに，第3.2.1項でも述べたように，悪意のあるWebサイトの同ドメイン内の異なるパスには別の悪意のあるWebサイトが存在する可能性が高いことがわかっている．そのため本研究では，そうしたWebサイトが互いにリンクを貼っているという仮説のもとにWebクローリングを行い，悪意のあるWebサイトからリンクが貼られているWebサイトのURLを収集している．本実験では，こうして集めたURLのうち，マルウェアを配布するWebサイトのものであるURLがいくつ含まれているかを検証する．Webクローリングを行い，集めたURLにVirtualBox上でアクセスを行い，マルウェア検体を収集できるかを確認し，本実験の実験結果として示す．

6.1.2 実験環境

検証実験の実験環境について述べる．表 6.1 に実験環境をまとめて示す．

表 6.1: Web クローリングの有効性の実験環境

要素名	利用環境	備考
OS	Ubuntu 11.04(32bit)	実機，Web クローリングを行う
	Windows XP SP なし (32bit)	VirtualBox，収集した URL にアクセス
SeedURL	173 個	Malware Domain List 掲載のもの
期間	2012.12.3∼2012.12.16	

6.1.3 実験結果

前項の実験環境において，Web クローリングを行い収集した Web サイトの URL にどの程度悪意のある Web サイトのものが含まれているのか検証した．以下の表 6.2 にて実験結果について述べる．

表 6.2: Web クローリングの有効性検証実験結果

	URL 数	収集した検体数
Seed となる URL	173 個	2 個
Web クローリングによって新たに収集した URL	31 個	4 個

なお，本実験は 2012 年 12 月 18 日に行った．そのため，すでに消滅している Web サイトもいくつか存在した．そのため，悪意のある Web サイトの発見と同時にクローリングを行うことで，より多くの検体を収集できると推測される．この結果についての評価と議論は第 7.1 節に後述する．

6.2 独自判断基準に基づく悪意のある Web サイト判定実験

本論文にて提案する手法に基づいた独自の判定基準を用いて様々な Web サイトの URL を判定し，悪意のある Web サイトと正常な Web サイトを正しく判定することができるかを検証する実験を行った．以下にこの実験の概要と実験環境を提示し，その結果を記述する．

6.2.1 実験概要

第 5.2.3 項にて述べたように，収集したすべての URL へのアクセスを行うことは，効率的とは言い難い．そのため，悪意のある Web サイトの特徴を抽出し，判定基準を作成した．その判定基準を学習データとした機械学習を行い，悪意のある Web サイトの可能性が高い Web サイトを判断する．そして，悪意があると判断された Web サイトから優先的にアクセスを行う．本実験では，その基準を用いた判定を行う際の誤検出，検出漏れの割合について検査し，その結果を示す．誤検出の割合を調査するため，正常な Web サイトに本研究にて提案する判定基準を当てはめ実験を行う．また，検出漏れの割合を調査するため，悪意のある Web サイトに本研究にて提案する判定基準を当てはめ実験を行う．検証手法は交差検証を用いて行った．正常なサイトは，Alexa[27] が提供する世界での閲覧数トップ 500 のサイト及び，Google[28] が提供する世界での閲覧数トップ 1000 のサイト [29] を使用する．なお，これらの 2 つのサイトでの情報にて重複する Web サイトは除いている．悪意のある Web サイトは，Malware Domain List 及び Malware Black List にて掲載されている URL を使用する．なお，本実験にて使用する判定基準とは第 4.3.1 項にて述べたものである．

6.2.2 実験環境

判定実験の実験環境について述べる．表 6.3 に実験環境をまとめて示す．

表 6.3: 悪意のある Web サイトの判定実験環境

要素名	利用環境	備考
OS	Ubuntu 11.04(32bit)	実機
	Windows XP SP なし (32bit)	VirtualBox
Alexa	500 の URL	正常な Web サイトの URL
Google ad planner	1000 の URL	正常な Web サイトの URL
Malware Domain List	2325 の URL	悪意のある Web サイトの URL
Malware Black List	2009 の URL	悪意のある Web サイトの URL

6.2.3 検証手法

判断基準の精度を検証するための手法として交差検証を利用した．交差検定とは，標本データ群を分割しその一部を解析して，残る部分で解析のテストを行い，解析自身の妥当性の検証・確認に当てる手法である．データの解析がどれだけ本当に母集団に対処できるかを良い近似で検証・確認するための手法である．本研究では，K-分割交差検証法と呼ばれる方法を用いて検証を行った．

K-分割交差検証法

K-分割交差検証法とは，標本データ群を K 個に分割し検証を行う方法である．本研究では図 6.1 の通り，悪意のある Web サイトと正常な Web サイトをそれぞれ標本データ群とした．

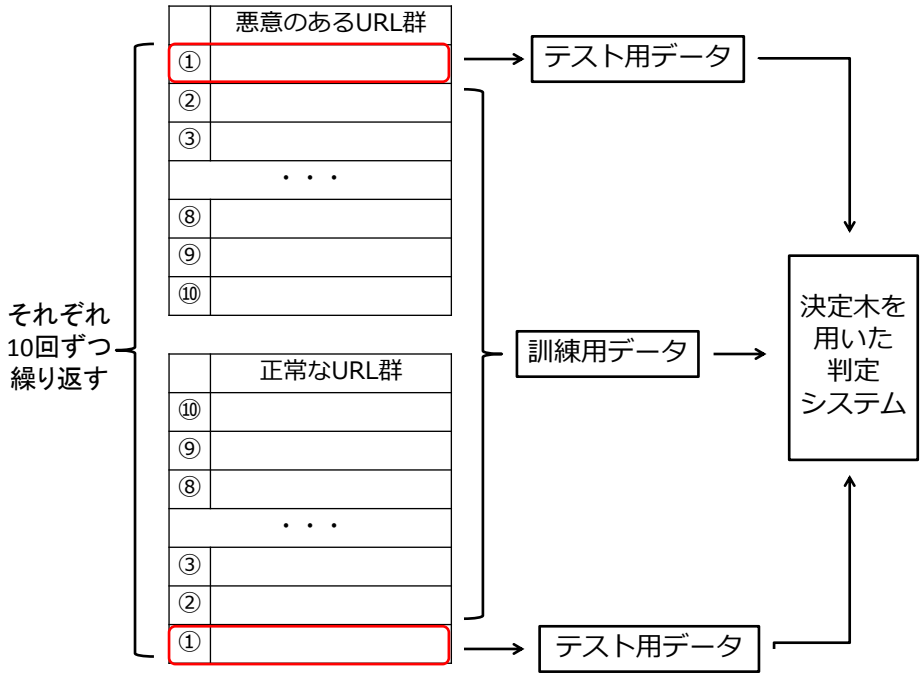


図 6.1: 本研究における 10 分割交差検証

そして，それぞれを 10 個に分割し検証を行った．本研究では 10 個に分割した標本データ群の内の 1 個をテスト事例とし，残る 9 個を訓練事例とした．その際に，悪意のある Web サイトの標本データ群は，Malware Domain List 及び Malware Black List の URL リストをそれぞれ 10 分割し，分割後テスト事例と訓練事例にそれぞれ結合した．また，正常な Web サイトの標本データ群は，Alexa 及び Google ad planner の URL リストを結合し，重複する URL を削除した後 10 分割した．そして，悪意のある Web サイトと正常な Web サイトの標本データ群に対してそれぞれ 10 回ずつ検証を行った．その結果を平均し，実験結果について考察を行った．第 6.2.4 項にてその結果について述べる．

6.2.4 実験結果

前項の実験環境において，本研究にて使用する判定基準の正確性について検証した．以下の表 6.4 にて，悪意のある Web サイトの URL をテスト事例とした実験結果について述べる．また，表 6.5 にて，正常な Web サイトの URL をテスト事例とした実験結果について述べる．

表 6.4: 悪意のある Web サイトの判定実験結果

	FN 数	TN 数	URL 総数	FN 率	TN 率
1	47	230	432	10.88%	53.24%
2	30	298	432	6.94%	68.98%
3	43	306	432	9.95%	70.83%
4	19	355	432	4.40%	82.18%
5	8	353	432	1.85%	81.71%
6	9	351	432	2.08%	81.25%
7	10	333	432	2.31%	77.08%
8	12	336	432	2.78%	77.78%
9	21	322	432	4.86%	74.54%
10	23	294	446	5.16%	65.92%
平均・合計	222	3,178	4,334	5.12%	73.33%

表 6.5: 正常な Web サイトの判定実験結果

	FP 数	TP 数	URL 総数	FP 率	TP 率
1	19	61	121	15.70%	50.41%
2	11	65	121	9.09%	53.72%
3	15	63	121	12.40%	52.07%
4	35	45	121	28.93%	37.19%
5	21	62	121	17.36%	51.24%
6	23	52	121	19.01%	42.98%
7	14	62	121	11.57%	51.24%
8	13	54	121	10.74%	44.63%
9	28	57	121	23.14%	47.11%
10	19	63	127	15.70%	52.07%
平均・合計	198	584	1,216	16.28%	48.03%

実験を行った結果，悪意のある Web サイトを平均 73.33%の割合で正しく判定できた．また，悪意のある Web サイトを正常な Web サイトであると判定してしまう検知漏れの割合は平均 5.12%であった．さらに，正常な Web サイトを悪意のある Web サイトであると判定してしまう誤検知の割合は平均 16.28%であった．この結果についての評価と議論は第 7.2 節に後述する．

6.3 マルウェア検体の収集実験

実装したシステムを実際に動かすことで，Malware Domain List 及び Malware Black List に掲載されている URL を Seed とした場合では，最終的にどのくらいのマルウェア検体を収集することができるのかを検証した．以下にこの実験の概要と実験環境を提示し，その結果を記述する．

6.3.1 実験概要

本実験では，本論文にて提案する手法を用いて設計・実装を行ったマルウェア収集システムを使用し，マルウェアの収集を行う．まず，Malware Domain List 及び Malware Black List にスクレイピングを行い，掲載されている URL を収集した．さらに，その収集した URL を Seed として Web クローリングを行うことにより，悪意のある Web サイトの可能性のある Web サイトの URL を収集した．その URL に対して第 4.3.1 項にて述べた判定基準を用いることで，悪意のある Web サイトの可能性が高いサイトの URL を抽出した．最後に，抽出した URL に VirtualBox を使用してアクセスすることで，マルウェアを収集した．なお，VirtualBox を使用して Web サイトにアクセスを行う際に，ホスト OS にてパケットキャプチャのツールである Wireshark[30] を使用した．これにより，マルウェアの行う通信の取得及び，目視にて確認できない部分で感染活動を行うマルウェアの検体を収集することが可能となる．なお，本研究では収集したファイルは VirusTotal[18] を用いてマルウェアか正常なファイルかを判断している．以下に，本実験を行った収集結果を示す．

6.3.2 実験環境

判定実験の実験環境について述べる．表 6.6 に実験環境をまとめて示す．

表 6.6: マルウェア検体の収集実験環境

要素名	利用環境	備考
OS	Ubuntu 11.04(32bit)	実機，Web クローリングを行う
	Windows XP SP なし (32bit)	VirtualBox，収集した URL にアクセス

6.3.3 実験結果

前項の実験環境において，本研究の手法に基づいたマルウェア検体の収集がどの程度可能であったか検証した．以下の表 6.7 にて実験結果について述べる．

表 6.7: マルウェア検体の収集実験結果

収集期間	総 URL 数	検体数（検体種類）
2013.1.8~2013.1.15	309URL	13 検体（9 種類）

2013 年の 1 月 8 日から 2013 年の 1 月 15 日までの約 1 週間の間に実験を行った．その結果，309 の URL から 13 個のマルウェア検体を収集することが可能であった．また，13 個の検体それぞれの MD5 ハッシュ値を元に区別した結果，9 種類のマルウェア検体であることが分かった．本論文では，収集したファイルを VirusTotal にアップロードしマルウェアかグッドウェアかどうかを判断する．その際に，VirusTotal にアップロードされた記録の無いファイルがいくつか発見された．この結果についての評価と議論は第 7.3 節に後述する．

6.4 まとめ

本章では，第 5 章にて述べた提案手法に基づいて設計，実装したシステムを用いて行った 3 種類の実験とその結果について述べた．まずはじめに，本システムにおける Web クローリングの有効性の検証実験について述べた．ここでは，悪意のある Web サイトの URL を SeedURL として Web クローリングを行い，収集した Web サイトの URL にどの程度悪意のある Web サイトが含まれているのか検証した．次に，独自の判断基準に基づいた悪意のある Web サイトの判定実験について述べた．ここでは，本論文にて提案する手法に基づいた独自の判定基準を用いて様々な Web サイトの URL を判定し，悪意のある Web サイトと正常な Web サイトを正しく判定できるかを交差検定し検証した．最後に，本システムを使用したマルウェア検体の収集実験について述べた．ここでは，実装したシステムを実際に動かすことで，Malware Domain List 及び Malware Black List に掲載されている URL を Seed とした場合では，最終的にどのくらいのマルウェア検体を収集することができるのかを検証した．第 7 章では，本研究にて実装したシステムがもたらす実験結果についての評価を行い，議論を展開する．

第7章　評価

本章では，第6章にて述べた，本研究にて実装したシステムがもたらす実験結果について，Webクローリングの有効性，判定部分の精度，収集検体数の3つの観点から評価を行い，議論を展開する．

7.1　Webクローリングの有効性評価

ここでは，悪意のあるWebサイトのURLを集める際のWebクローリングの有効性について検証した第6.1節の実験結果について評価と議論を行う．実験では，173個の悪意のあるWebサイトからそれぞれWebクローリングを行い，その有効性について検証した．その結果，31個の新しいWebサイトのURLを得ることができ，そのURLから2個の新しいマルウェア検体を収集することができた．悪意のあるWebサイトには，マルウェアを配布する目的で作成されたものと，正常なWebサイトが改ざんされマルウェア配布サイトへのリダイレクトを行うものが存在する．本実験により，マルウェアを配布する目的で作成された悪意のあるWebサイトは，他の悪意のあるWebサイトからリンクが貼られている場合があることを証明することができた．そのため，Webクローリングと，本研究独自の悪意のあるWebサイト判定方法を組み合わせることで，効率的にマルウェア検体を収集できると推測できる．

7.2　判定部分の精度評価

ここでは，Webクローリングを行い収集したURLから悪意のあるWebサイトを抽出するための，URLの判定部分の精度について検証した第6.2節の実験結果について評価と議論を行う．実験では，決定木という機械学習方法を用いて判定を行った．4,334個の悪意のあるWebサイトと1,216個の正常なWebサイトをそれぞれ10分割し，1つをテスト用データ，9つを訓練用データとして10分割交差検定によって判定を行った．その結果，TNが平均73.33%の割合となった．このことから，本研究にて使用した入力データを用いて判定を行うことで，高い精度で悪意のあるWebサイトを特定できるということが分かっている．一方で，正常なWebサイトを正しく判定することができた割合は十分とは言えない精度であった．しかしこれは，悪意のあるWebサイトに対して訓練用のデータが少ないことが理由であると想定される．さらに，他の研究と比較した際に，本研究にて使用した訓練用のデータは必ずしも多いとは言えない．そのため，訓練用データの数を増やすことで，より正確に悪意のあるWebサイトおよび正常なサイトを判定することがで

きると想定される．

また，FN の割合は平均 5.12%であった．そして，FP の割合は平均 16.28%であった．第 3 章にて述べた，C.Seifert らによる Identication of Malicious Web Pages with Static Heuristics では，46.15%の FN と 5.88%の FP が生じている．この研究では，5,678 個の悪意のある Web サイトと 16,006 個の正常なサイトを訓練用のデータとし，61,000 個の URL を判定している．表 7.1 にて先行研究による悪意のある Web サイトの判定精度と本研究によるものを比較する．本研究では，FN を防ぐことができる割合で先行研究を大きく上回った．FP の割合では先行研究に対して劣る結果となっているが，マルウェア解析者や研究者のために検体を収集するということが本研究における目的であるため，数多くのマルウェア検体を効率的に収集するためには，FN の割合を下げることが必要である．また，本研究における FP の割合は 16.28%と低く，悪意のある Web サイトを巡回する際の妨げにはならないと想定される．これより，本研究にて提案する判定基準を用いた悪意のある Web サイトの判定方法は有効であると言うことができる．

表 7.1: 先行研究との精度比較

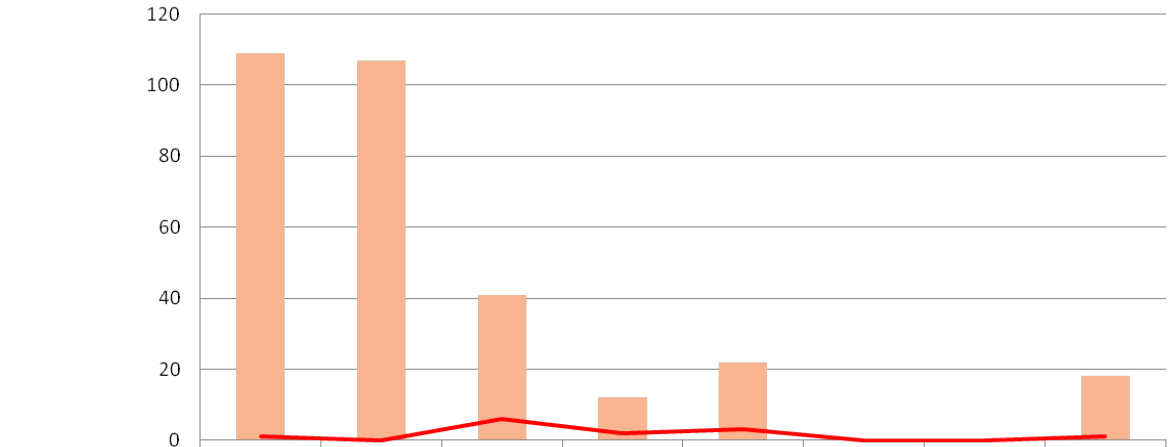
	FN 率	FP 率	TN 率
C.Seifert らの論文	46.15%	5.88%	記載なし
寺田 剛陽らの論文	12.3%	17.1%	85.1%
本論文	5.12%	16.28%	73.3%

7.3 収集検体数評価

ここでは，本論文にて提案・実装したシステムを実際に使用してどのくらいのマルウェア検体を収集することができたのかという第 6.3 節の結果について評価と議論を行う．実験では，2013 年の 1 月 8 日から 2013 年の 1 月 15 日にかけてシステムを動作させ，マルウェア検体を収集している．その結果 309 個の URL を収集することができ，そこから 13 個のマルウェア検体を収集することができた．また，13 個のマルウェア検体それぞれの MD5 ハッシュ値を取得し区別したところ 9 種類のマルウェア検体に区別された．

第 3 章にて述べた，青木らによる能動的攻撃と受動的攻撃に関する調査及び考察では，悪意のある Web サイトのものであると検知した 3,408 の URL から 9,533 個のマルウェア検体を収集している．しかし，SHA1 のハッシュ値を元に区別した結果，9,533 個のマルウェア検体は 136 種類に区別できることがわかっている．この論文における検知した URL と種類数について，URL 数に対する種類数の割合は 3.99%となっている．本論文における URL 数に対する種類数の割合は，2.91%であり，青木らの論文との差異はほとんどないことが分かる．そのため，総 URL 数を増加させることで，より効果的かつ効率的にマルウェア検体を収集することができることが推測される．

また，収集したファイルを VirusTotal にアップロードする際に，これまで VirusTotal に



	1月8日	9日	10日	11日	12日	13日	14日	15日
Seed URL	109	107	41	12	22	0	0	18
マルウェア検体	1	0	6	2	3	0	0	1

図 7.1: 日別収集検体数

アップロードされた記録のないファイルをいくつか発見することができた．さらに，本研究では定期的に Malware Domain List 及び Malware Black List を巡回している．そのため，悪意のある Web サイトの情報が更新されてから時間を置かずに情報を取得することが可能である．先行研究では，ウイルス対策ベンダなどから悪意のある Web サイトの URL リストの提供を受け，研究を行う場合が多くみられる．その場合は，まとめて URL のリストを受け取るため，取得した際にリストに記載されている URL がすべて最新のものであるとは限らない．本論文の手法では，そうした先行研究と比較すると労力，コストなどの面により，収集できる総 URL 数では劣ってしまう．しかし，常に新しい情報を取得し，その情報を元に Web サイトを巡回することができる点においては本研究が優れていると言える．こうしたことから，本論文のシステムではまだ一般に出回っておらず，被害をもたらしていない新規のマルウェア検体を発見・収集する可能性があるということが言える．

7.4 まとめ

本章では，第 6 章で述べた，本研究にて実装したシステムがもたらす実験結果について評価を行い，議論を展開した．まずはじめに，Web クローリングの有効性の評価を行った．この実験の結果から，悪意のある Web サイトは他の悪意のある Web サイトからリンクが貼られている場合があるということを証明でき，Web クローリングの有効性を証明することができた．次に，判定部分の精度の評価を行った．本論文では，悪意のある Web サイトを正常な Web サイトであるとして FN を起こしてしまう確率を 5.12%まで下げることができた．この結果は，先行研究と比較しても低い値であり，より多くのマルウェア検体を効率よく収集するという目的の中では必要なことである．最後に，収集検体数の評

価を行った．本論文では，309 個の URL から 9 種類のマルウェア検体を収集でき，URL 数に対する種類数の割合が 2.91%となった．この値は，先行研究と比較しても差異はほぼ見られない．しかし，URL の総数が少ないことは問題であり，総 URL 数を増加させることでより効果的効率的にマルウェア検体を収集することができるということが推測される．第 8 章では，本論文の全体についてをまとめ，達成できた目的について述べる．そして，本論文の発展を実現するために今後の展望を述べる．

第8章　結論

本章では，本論文の全体についてをまとめ，本論文の第1章にて述べた目的の中で達成したことについて述べる．そして，本論文の発展を実現するために今後の展望を述べる．

8.1　まとめ

本論文の目的は，効率的に悪意のあるWebサイトを巡回することで，効率的にマルウェア検体を収集することである．これにより，マルウェアの解析や検知を行う研究者にマルウェア検体の情報を提供することができる．また，より効果的な対策を迅速にとることができるようになることが期待される．さらにその実現のために，悪意のあるWebサイトに優先的にアクセスし，効率的にマルウェア検体を収集することができるようにすることが必要であると言える．そこで，本論文では上記のような目的を達成するための手法を提示した．まず，悪意のあるWebサイトの特徴を抽出し，機械学習のアルゴリズムである決定木を用いたシステムに訓練データとして学習させる．その結果，悪意のあるWebサイトの可能性が高いと判断されたWebサイトから優先的にアクセスする．そうして，正常なWebサイトにアクセスする時間を短縮することで，効率よくマルウェア検体を収集することができる．この手法を基にして，悪意のあるWebサイトに優先的にアクセスし，効率的にマルウェア検体を収集するシステムを設計・実装した．そして，このシステムを用いて悪意のあるWebサイトを判別することで，手法の有効性を検証した．判定によって悪意のあるWebサイトを正しく特定できた割合は73.33%であった．また，FPを16.28%に，検出漏れを5.12%に抑えることができた．より多くのマルウェア検体を収集するという目的の中では，FPの割合を低くしながらも検出漏れを可能な限り低くすることが必要である．本論文での検出漏れの値は，先行研究と比較しても大幅に抑えられた結果となっている．さらに，本システムを用いて2013年の1月8日から2013年の1月15日までの約1週間の間マルウェア検体を収集した．その結果，309個のURLを収集することができ，そこから13個（9種類）のマルウェア検体を収集することができた．本論文におけるURL数に対する種類数の割合は，2.91%であり，先行研究との差異はほとんどないことが分かる．そのため，総URL数を増加させることで，より効果的効率的にマルウェア検体を収集することができるということが推測される．本論文によって，対策の立案に先立ち，効率的にマルウェア検体を収集し，マルウェアの解析や検知を行う研究者にマルウェア検体の情報を提供することが可能となった．これにより，解析や検知を行う研究者は，提供された事前情報に基づいて解析および対策の立案を進めることができるため，効果的かつ効率的に作業を進めることができるようになるという可能性を提示した．

8.2 今後の展望

本論文全体のまとめを受け，本論文の発展のため，今後の展望を述べる．今後の展望については，判定部分とマルウェア検体収集部分の 2 つに分けて述べる．

8.2.1 判定部分の精度

ここでは，判定部分の精度について今後の展望を述べる．現在，悪意のある Web サイトを 73.33%の割合で特定することができ，FN を 5.12%に抑えることが可能である．しかし，決定木を用いたシステムに学習させている訓練データのうち，悪意のある Web サイトの URL が 4,334 個ある一方で，正常な Web サイトは 1,216 個しか存在しない．そのため，正常な Web サイトは 48.03%の割合でしか特定することができない．正常な Web サイトのデータを悪意のある Web サイトと同等の個数まで増やすことで，FP の割合も 16.28%から減少することが見込まれる．これによって，情報の信頼性の向上に繋がると期待される．さらに，学習させる訓練データに，より多くの悪意のある Web サイトに多くみられる特徴のデータを加えることで悪意のある Web サイトの検出率をさらに向上させることが可能であると推測する．

8.2.2 マルウェア検体収集効率

ここでは，本論文にて提案・実装したシステムを用いたマルウェア検体の収集効率について今後の展望を述べる．現在，約 1 週間で 309 個の URL を取得し，13 個（9 種類）のマルウェア検体を収集することができている．総 URL 数におけるマルウェア検体種類数の割合は 2.91%と他の研究と差異は少ないものの，やはりウイルス対策ソフトウェアベンダ等の企業の提供する悪意のある Web サイトのリストと比較すると総 URL 数が少ないということが分かる．総 URL 数を増加することで悪意のある Web サイトの URL 数を増やすことができ，収集可能なマルウェア検体数が増えることが判明している．そのため，Web クローリング部分を改良し，Seed となる URL を増やすことで総 URL 数を増やすことができると推測する．まず，Google 等の検索エンジンにて「malware」や「computer virus」といった様々な単語を検索した結果を取得する．次に，悪意のある Web サイトに含まれるコンテンツの特徴や，Malware Domain List 及び Malware Black List のような，悪意のある Web サイトの情報を複数記載する Web サイトの特徴を抽出する．こうした結果を組み合わせることにより，Malware Domain List や Malware Black List のような Web サイト及び，こうした Web サイトに記載されていない悪意のある Web サイトを複数見つけることができると推測される．

上記の 2 つを改良することにより，本システムにおける判定の精度や収集効率を向上できることが期待できる．

謝辞

本論文の作成にあたり，ご指導頂いた慶應義塾大学環境情報学部学部長 村井 純博士，同学部教授 徳田 英幸博士，同学部教授 中村 修博士，同学部准教授 楠本 博之博士，同学部准教授 高汐 一紀博士，同学部准教授 三次 仁博士，同学部准教授 植原 啓介博士，同学部 専任講師 中澤 仁博士，同学部准教授 Rodney D.Van Meter III 博士，同学部教授 武田 圭史博士，同大学政策・メディア研究科特任講師 斉藤 賢爾博士，同大学政策・メディア研究科特別研究講師 佐藤 雅明博士に感謝致します．特に武田圭史博士は，研究で行き詰まる私に対して非常に根気強く指導して下さり，常に新しいアイディアと研究手法で私を導いていただくことで何度も私に新しい視点や手本を見せていただきました．本当にありがとうございました．

そして，本研究を進めていく上で様々な励ましと助言，お手伝いをいただきました，村井研究室卒業生である水谷 正慶氏，金井 瑛氏，上原 雄貴氏，重松 邦彦氏，梅田 昴翔氏，福岡 英哲氏，相見 眞男氏，Doan Viet Tung 氏，Vu Xuan Duong 氏，Pham Van Hung 氏，吉原 洋樹氏に感謝致します．

慶應義塾大学政策・メディア研究科修士課程，碓井 利宣氏，関根 冬輝氏，山本 知典氏に感謝致します．特に碓井 利宣氏には，平日休日問わず 24 時間体制で自らの健康も顧みず親身に相談に乗っていただき，研究の方向性や実装のあらゆる面，生活の面でも面倒を見ていただきました．氏なしでは卒論執筆はもちろんのこと，充実した研究室生活を送ることはできませんでした．本当に感謝いたします．

研究室で苦楽を共にした有馬 怜文氏，大矢 崇央氏，Nguyen Anh Tien 氏，鴻野 弘明氏，小松 真氏，中島 明日香氏，由井 卓哉氏，露木 航平氏，中安 恒樹氏，川本 卓弥氏，廣田 一貴氏，Tran Ngoc Anh 氏に感謝致します．彼らと一緒に研究をすることでお互いを刺激しあい，より質の高い議論や研究をすることができました．特に，大野三津雄氏，恩田優氏，高岡賢二氏，三ツ木あかね氏を含めた武田圭史研究会の同級生には，研究室内での活動だけでなく，対外発表や対外活動，レクリエーションなど様々な場面でお世話になりました．この場を借りてお礼を述べさせていただきます．また，由井 卓哉氏には研究に行き詰った際に何度も的確なアドバイスをいただき，とてもお世話になりました．本当に感謝いたします．

卒論執筆をするにあたって私の荒んだ心を癒してくれた社長や専務たちに感謝いたします．彼らのおかげで心に余裕をもって卒論の執筆を行うことができたと確信しています．私の大学 4 年間の心の拠り所であった慶應ライナーズのメンバー全員に感謝いたします．体を動かすことで執筆のストレスから解放されたことも少なくなかったと思っています．特に，同級生には研究や卒論執筆について様々な相談をさせていただき，とてもお世話になりました．また，同サークルの藤井望氏に感謝したいと思います．彼女には概要部分の

英語執筆における添削も含め，様々な面でサポートしていただきました．この場をかりてお礼を述べさせていただきます．

最後に，大学入学からの 4 年間だけでなくこれまでの 22 年間をあらゆる面で支えてくれた父，吉原 雅と母，吉原 美也子，弟の吉原 拓郎に心から感謝いたします．

参考文献

[1] McAfee. http://www.mcafee.com/, 12 2012.

[2] McAfee Labs. Mcafee 脅威レポート:2012 年第 1 四半期, 2012.

[3] Trendmicro. http://jp.trendmicro.com/, 12 2012.

[4] M. Akiyama, T. Yagi, and M. Itoh. Searching structural neighborhood of malicious urls to improve blacklisting. *2011 IEEE/IPSJ International Symposium on Applications and the Internet*, pages 1–10, 2011.

[5] Facebook. http://www.facebook.com/, 12 2012.

[6] Seculert. http://www.seculert.com/, 12 2012.

[7] 独立行政法人 宇宙航空研究開発機構 JAXA. http://www.jaxa.jp/, 12 2012.

[8] ESET. Eset smart security 5. http://www.eset.com/us/, 12 2012.

[9] キャノン IT ソリューションズ株式会社. マルウェアランキング. http://canon-its.jp/product/eset/topics/malware.html, 9 2012.

[10] Nepenthes. http://nepenthes.mwcollect.org/, 3 2008.

[11] 谷本直人, 八木毅, 針生剛男, and 伊藤光恭. 複数のドメインに配置されたハニーポットを用いた web サイトへの攻撃の実態調査. **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**, 2010.

[12] 北村真一, 岩村誠, and 伊藤光恭. クライアント型ハニーポットによる悪意ある web サイトの検出について. **電子情報通信学会総合大会講演論文集**, (2):124, 2008.

[13] Mitsuaki Akiyama, Makoto Iwamura, Yuhei Kawakoya, Kazufumi Aoki, and Mitsutaka Itoh. Design and implementation of high interaction client honeypot for drive-by-download attacks. *IEICE Transactions*, pages 1131–1139, 2010.

[14] 株式会社フォーティーンフォティ技術研究所. http://www.fourteenforty.jp/, 1 2013.

[15] 株式会社フォーティーンフォティ技術研究所. Origma+ 製品概要. http://www.fourteenforty.jp/products/origma/, 1 2013.

[16] NTT 東日本. http://www.ntt-east.co.jp/, 11 2012.

[17] NTT コミュニケーションズ. Nttcommunications. http://www.ntt.com/, 11 2012.

[18] virustotal. https://www.virustotal.com/, 11 2012.

[19] 星澤裕二, 川守田和男, 太刀川剛, and 神薗雅紀. 自律型クライアントハニーポットの提案 (高度インシデント分析を支える要素技術, インターネットセキュリティ, 一般). **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**, 109(86):13–17, 2009.

[20] 青木一史, 川古谷裕平, 秋山満昭, 岩村誠, 針生剛男, and 伊藤光恭. 能動的攻撃と受動的攻撃に関する調査及び考察. **情報処理学会論文誌**, 50(9):2147–2162, 2009.

[21] Microsoft. http://www.microsoft.com/, 12 2012.

[22] Christian Seifert, Ian Welch, and Peter Komisarczuk. Identification of malicious web pages with static heuristics. *In Proceedings of the LCN Workshop on Network Security (WNS)*, 2008.

[23] 寺田剛陽, 古川忠延, 東角芳樹, and 鳥居悟. 検知を目指した不正リダイレクトの分析. **情報処理学会シンポジウム論文集**, pages 765–770, 2010.

[24] Niels Provos, Dean McNamee, Panayiotis Mavrommatis, Ke Wang, and Nagendra Modadugu. The ghost in the browser analysis of web-based malware. In *Proceedings of the first conference on First Workshop on Hot Topics in Understanding Botnets*, 2007.

[25] Malware Domain List. http://www.malwaredomainlist.com/, 12 2012.

[26] Malware Black List. http://www.malwareblacklist.com/, 12 2012.

[27] Alexa. http://www.alexa.com/topsites/global, 12 2012.

[28] google. http://www.google.com, 12 2012.

[29] google. The 1000 most-visited sites on the web. http://www.google.com/adplanner/static/top1000/, 12 2012.

[30] wireshark. http://www.wireshark.org/, 12 2012.