Computer Security Symposium 2018
22 - 25 October 2018

Drive-by-Download攻撃時の通信データの関連性分析

高田 真資[1]　高橋 健一[2,3]　川村 尚生[2,3]　菅原 一孔[2,3]

概要：Drive-by-Download 攻撃は，Web ユーザーに気づかれることなくマルウェアをダウンロードさせる攻撃であり，攻撃の際には正常な Web サイトへのアクセス時に見られない通信の特徴が存在する．これまでに通信の特徴を用いて攻撃の検知を試みる研究が複数行われており，我々は先行研究においてそれらの研究を調査し，検知に使われる通信の特徴を検知項目と定義した．本研究では，既に提案された検知手法の妥当性の裏付け，精度の高い新たな検知手法を見つけることを目的としており，そのために各検知項目について相関分析，アソシエーション分析を行い，その関連性について分析した．

キーワード：Drive-by-Download 攻撃，相関分析，アソシエーション分析

Analysis on relevance of communication data during Drive-by-Download attack

Masashi Takada[1]　Kenichi Takahashi[2,3]　Takao Kawamura[2,3]　Kazunori Sugahara[2,3]

Abstract: The Drive-by-Download attack makes Web users download a malware without being noticed. There is some characteristics of communication during the attack which are not seen when accessing a normal Web site. Several methods using the characteristics of communication have been studied for the detection of attacks. We have surveyed those detection methods and defined the characteristics of communication used for detection as detection items. In this research, we aim to evaluate the validity of the already proposed detection method and find a new detection method with high efficiency or high accuracy. We perform correlation analysis and association analysis for each detection item, and analyze correlation and relevance of the detection items.

Keywords: Drive-by-Download attack, correlation analysis, association analysis

1. はじめに

Web におけるマルウェアの感染経路として Drive-by-Download 攻撃が存在する．Drive-by-Download 攻撃は，Web サイトにアクセスした Web ユーザーが意図せずにマルウェアをダウンロードさせられる攻撃である．標的型攻撃においてターゲットをマルウェアに感染させる手段としても用いられる．攻撃者は，正規の Web サイトの改ざんや独自にサイトを作成することで入り口サイトを用意し，入り口サイトに訪れた Web ユーザーを悪性サイトにリダイレクトさせる．Web ユーザーは攻撃サイトにおいてユーザー PC に存在するソフトウェアの脆弱性を利用されマルウェア配布サイトからマルウェアをダウンロードさせられる．このように，マルウェアへの感染は入り口サイトから踏み台サイト，攻撃サイトを経て行われる．また，近年，Drive-by-Download 攻撃では Exploit Kit[1] と呼ばれる攻撃支援ツールが用いられる傾向にある．これは攻撃に必要な攻撃コードやマルウェア配布サイトなどがパッケージ化されたものであり，攻撃者の負担を減らし攻撃を容易にしている．Exploit Kit を使用した攻撃では，類似したシェルコードやサーバの設定が用いられるため，攻撃時の通信の特徴についても似たような傾向で出現すると考えられる．

[1] 鳥取大学大学院持続性社会創生科学研究科
Graduate School of Sustinability Science, Tottori University
[2] 鳥取大学大学院工学系研究科
Graduate School of Engineering, Tottori University
[3] 鳥取大学工学部附属クロス情報科学研究センター
Tottori University Cross-informatics Research Center

Drive-by-Download 攻撃では上で示したような通信が発生するため，その通信の特徴を利用することで攻撃の検知を行う手法が提案されている．

我々は先行研究において，攻撃の検知手法を調査し，検知に用いられる通信の特徴を検知項目として定義した．検知項目は，Web ページにアクセスした際に読み込まれるページの階層構造や，リダイレクトされた際のホストの遷移，ユーザーエージェントの変化などの特徴である．先行研究においては，Web ページにアクセスしてからの一連の通信の中で，これらの特徴が同時に出現した場合や時間的に近いタイミングで出現した場合などに悪性として検知を試みている．

そこで，本研究では良性・悪性通信データ中に出現した検知項目を調べ，その相関関係を分析した．また，各データにおける検知項目の出現の有無からアソシエーション分析を行い，相関ルールを抽出した．良性データ・悪性データから抽出した相関ルールについて傾向を分析した．

2. 関連研究

Drive-by-Download 攻撃により発生する通信から攻撃の検知を試みる研究が存在する．通信から検知を行う場合は URL や IP アドレス，HTTP ヘッダ，受信するコンテンツの内容などの情報を用いる．北野ら [2] の研究では Drive-by-Dwonload 攻撃時の通信中に Exploit Kit の種別によらず発生する定性的な特徴（受信するデータ量の遷移，ホストの遷移，ファイルの種類など）から検知を行う．酒井ら [3] の研究では HTTP レスポンスヘッダに PHP のバージョン情報が含まれるか，危険度の高いファイルを受信するかを用いて検知を行う．安藤ら [4] の研究では Web ページのリンクの深さと広がりという概念を用いて，Drive-by-Dwonload 攻撃で発生する不正なリダイレクトを検知する．その際に，リダイレクト時に通信先のホストが変化するか，危険度の高いファイルを受信するかといった情報も組み合わせて検知を行う．佐藤ら [5] の研究でも Web ページのリンク構造に着目しており，さらに，URL から得られる PageRank やドメイン年齢といった情報を用いて検知を行う．また，工藤ら [6] の研究では解析時のコストを考慮し，より単純なヘッダ情報の組み合わせから不正なリダイレクトの検知を行う．寺田ら [7] の研究では，脆弱性に対し攻撃を行うファイルと，攻撃の結果，ダウンロードされる実行ファイルが通信中に出現するかどうかから検知を行う．

我々は先行研究において，上であげた研究で用いられる通信の特徴を検知項目としてまとめた．本研究では，それぞれの検知項目の関連性を調べるために相関分析，アソシエーション分析を行った．

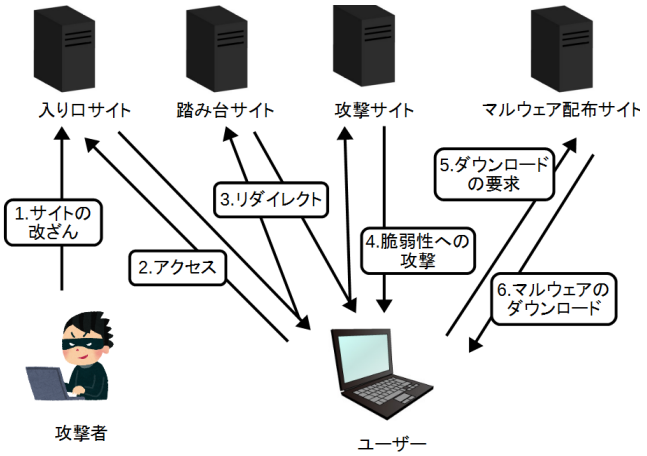

図 1 Drive-by-Download 攻撃の流れ
Fig. 1 Flow of Drive-by-Download Attack

3. Drive-by-Download 攻撃

Drive-by-Download 攻撃は主に，入り口サイト，踏み台サイト，攻撃サイト，マルウェア配布サイトからなる．攻撃の流れを図 **1** に示す．まず攻撃者は入り口サイトを作成する．これは攻撃者が新たにサイトを作成したり，一般の Web サイトを改ざんすることで行う．Web ユーザーが入り口サイトにアクセスすると，自動的にリダイレクトが発生し，踏み台サイトを経由して攻撃サイトへ誘導させられる．この時，踏み台サイトでのリダイレクトが複数回生じる場合がある，攻撃サイトではユーザーの PC 環境に応じて，その脆弱性を利用した攻撃が行われる．その結果，Web ユーザーはマルウェア配布サイトからマルウェアをダウンロードさせられる．

Drive-by-Download 攻撃により発生した通信には特徴があり，先行研究では攻撃の検知のために以下の特徴が用いられていた．

A：Web 階層の深さ

Drive-by-Download 攻撃では，リダイレクト時に複数の踏み台サイトを経由する．そのため，深い Web 階層で読み込まれる Web ページの信頼度は低いと考えられる．

Web 階層とは Web ページにアクセスした際に，そのページが別のページを読み込むことによって生じる Web ページの階層的な構造である．ある Web ページ A にアクセスした際，A から別の Web ページ B を読み込み，B からさらに別の Web ページ C を読み込む，というように階層的にページの読み込みが繰り返される場合がある．この時，始めにアクセスした Web ページ A の階層を 1，読み込まれる Web ページ B の階層を 2，さらに先の Web ページ C の階層を 3 というように Web 階層をカウントする．

B：ホストの遷移

入り口サイトはユーザーのアクセスを促すために一般の Web サイトを改ざんして用意されるが，攻撃サイトは攻撃

のため攻撃者によって用意されることが多い．そのため，Web ユーザーが入り口サイトから攻撃サイトにリダイレクトされる際にホストの遷移が発生する．

C：ファイルの種類

Exploit Kit ではソフトウェアの脆弱性を用いて攻撃を行う．中でも Java，Flash の脆弱性が利用されることが多く.jar や.swf を読み込む通信は悪性の可能性が高いといえる．また，実行形式のファイルはマルウェア本体である可能性があり，危険度が高い．上記を踏まえ，通信中に現れた特定の種類のファイルを検知する．ファイルの種類は，リクエストした URL 中の拡張子，HTTP の Content-Type ヘッダに格納される MIME タイプ，ファイルのマジックナンバーから確認する事ができる．

D：MIME タイプとマジックナンバーの整合性

マルウェアのダウンロード時には，ファイルがマルウェアであることを偽装するために，実際にダウンロードされるファイルの種類とは異なる MIME タイプが Content-Type ヘッダに設定されることがある．当然，正常な通信の場合は，MIME タイプは受信するファイルと対応したものになる．そのため，Content-Type に設定された MIME タイプとマジックナンバーの整合性がとれていない場合は，悪性である可能性が高いと考えることができる．

E：X-Powered-By ヘッダの値

X-Powered-By ヘッダには通信先のサーバソフトウェアの情報が設定される．サーバソフトウェアが Apache の場合，X-Powered-By ヘッダにはサーバで使用する PHP のバージョン情報が設定される．Exploit Kit を用いた Drive-by-Download 攻撃では，X-Powered-By ヘッダに古い PHP バージョンが記載されることが多い．

F：受信データ量の遷移

Drive-by-Download 攻撃では，リダイレクト時，攻撃時などの段階ごとにユーザーが受信するデータの大きさが変化する．入り口サイトへのアクセス時，リダイレクト時には数百〜数千バイトほどの大きさだが，攻撃ファイルのダウンロード時には数十〜数百キロバイトに上昇する．そのため，このような受信データ量の変化がみられたら攻撃の可能性が高いと考えれる．しかし，Web では複数の画像，動画ファイルを読み込むことも多く，必ずしも悪性な通信でのみみられる特徴ではない．

G：UserAgent の変化

JRE の脆弱性を利用した Exploit Kit では攻撃サイトにアクセスする際に，Java アプリケーションにより通信が行われるため，UserAgent が Java となる．また，マルウェアが C & C サーバと通信する際には，ユーザーのブラウザ環境とは異なる UserAgent が出現する．そのため，通信中に UserAgent が変化した場合は悪性の可能性が高いと考えられる．

4. 分析データ

4.1 悪性データ

悪性データとして D3M データセット [8] に含まれる攻撃通信データを使用した．D3M(Drive-by-Download Data by Marionette) データセットは NTT セキュアプラットフォーム研究所の Web クライアント型ハニーポット (Marionette) により収集された Web 感染型マルウェアの観測データ群である．D3M データセットには，ブラックリストに登録された URL をハニーポットが巡回した際にキャプチャした攻撃通信データ，巡回した URL が収録されている．攻撃通信データは pcap 形式のファイルであり，1 つのファイル内に複数種類の URL へアクセスした際の通信データが収録されている．

4.2 良性データ

良性データとして，Alexa によるランキング [9] 上位の Web サイトや官公庁，地方自治体，大学が運営する Web サイトを良性と判断して独自に収集を行った．ここで収集した Web サイトは悪性である可能性もあるが，その確率が低いと考えられるため，すべて良性として収集した．収集では，良性サイトと判断したサイトにアクセスした際の通信データを Wireshark によりキャプチャし，pcap 形式で保存した．

4.3 検知項目の出現数の抽出

悪性サイト，良性サイトにアクセスした際の pcap データから検知項目の出現数を抽出する．その際，HTTP の Referer ヘッダ，Location ヘッダの情報をもとにリダイレクト元，リダイレクト先を結び付け，最初に悪性 (良性) サイトにアクセスする際の HTTP リクエストからそれ以上リダイレクトが発生しなくなるまでの一連の通信の流れをリダイレクトチェイン（**図 2**）とした．良性通信データからは 765 個，悪性通信データからは 983 個のリダイレクトチェインを得られた．検知項目はこのリダイレクトチェインごとに抽出した．

Web 階層については 5 階層以上の Web ページについてのみ調べた．ホストの遷移については，リダイレクトチェイン中で通信先のホストが変わったかどうかを特徴とした．ファイルの種類については，exe ファイル，swf ファイル，jar ファイル，pdf ファイル，js ファイルを特徴的なファイルとした．ファイルの種類の判断には以下の情報を用いた．

(1) HTTP リクエストの URL に含まれる拡張子
(2) Content-Type ヘッダに設定された MIME タイプ
(3) 受信したファイルのマジックナンバー

exe ファイルについては，(1)，(2)，(3) の情報を用いて，

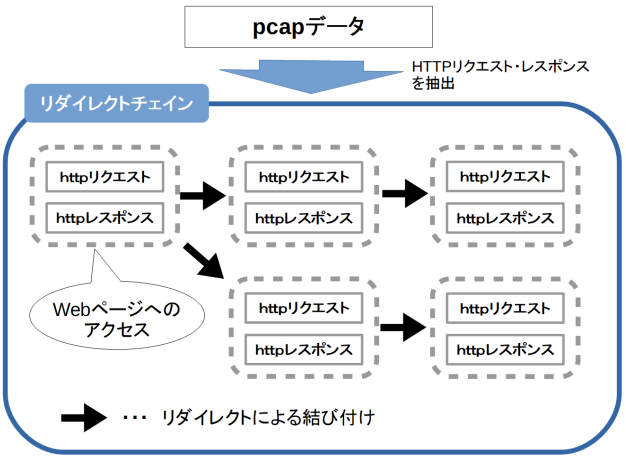

図 2　リダイレクトチェイン
Fig. 2　Redirect Chain

exe 以外のファイルは（1），（2）の情報を用いて判断した．MIME タイプとマジックナンバーの整合性については，実行ファイルが偽装されている場合についてのみ調べた．X-Powered-By ヘッダの値については，ヘッダ内に含まれる PHP のバージョン情報が 5.4.0 より古いバージョンのものだけを抽出した．受信データ量については，HTTP 応答で 100 キロバイト以上のデータを受け取った場合を特徴的な通信として利用した．UserAgent の変化については，リダイレクトチェイン中に UserAgent に変化があったかどうかを特徴とした．

5.　相関分析

リダイレクトチェイン中に出現した検知項目の出現数から相関分析を行った．相関分析は 2 変数間の関係を分析する方法であり，今回の分析ではピアソンの積立相関分析により検知項目同士に直線的な比例関係があるかを分析した．ピアソンの積立相関分析では以下の式で 2 変数 x，y の相関係数 r を求める．

$$r = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2}}$$

上式の分母はサンプル数 N のデータから得られた x の標準偏差と y の標準偏差を掛け合わせてたものであり，分子は x と y の共分散である．x と y の共分散は x と y の関係を表す数値となる．

相関係数 r は一般的に**表 1** のような意味を持つ．上で示した式の x，y それぞれに検知項目の各リダイレクトチェインでの出現数をあてはめ，相関係数を求めた．全ての検知項目の組み合わせについて相関係数をもとに相関のある組み合わせを調べた．

表 1　相関係数
Table 1　Correlation Coefficient

$0.7 < \|r\|$	強い相関がある
$0.4 < \|r\| \leq 0.7$	相関がある
$0.2 < \|r\| \leq 0.4$	弱い相関がある
$\|r\| \leq 0.2$	相関がない

5.1　相関のある組み合わせ

相関係数が 0.2 以上となった組み合わせについて**表 2** に示す．良性データでは 10 個，悪性データでは 19 個見られた．

5.1.1　良性データから得られた相関のある組み合わせ

js ファイルを示す特徴とホストの遷移の相関（bc3，bc6），js ファイルを示す特徴と受信データ量の相関（bc8，bc9）が見られた．現在は，多くの Web サイトで JavaScript が使用されており，Web ページのデータ量も増大する傾向にある．また，SNS へのリンクや Web 広告，アクセス解析サービスの利用などで別のホストのファイルを読み込むことも多く，ホストの遷移も発生しやすい．その結果，js ファイル，ホストの遷移，受信データ量といった特徴は良性通信中にも頻繁に見られており，それらには bc3，bc6，bc8，bc9 のように相関があると考えれる．また，ホストの遷移と Web 階層 5 以上の相関（bc1），拡張子.js と Web 階層 5 以上の相関（bc2）は良性でのみ見られた．これについて，Web 階層が 5 以上の Web ページが多く出現した場合，リダイレクトが複数生じており，リダイレクトチェイン中に現れるファイルの数も多くなるため，結果的に js ファイルやホストの遷移が出現する確率も上がる．それにより，bc1 や bc2 には相関がみられたと推測できる．ただし，bc1，bc2 のような相関関係は悪性データからは見られなかった．

exe を示す MIME タイプと拡張子.js，ホストの遷移の相関（bc5，bc7）も見られた．ここでの exe を示す MIME タイプは application/octet-stream，application/x-msdownload，application/x-download，application/x-msdos-program の 4 つである．良性データ中に出現した exe を示す MIME タイプは全て application/octet-stream を判断したものであった．そのため，実際のファイルの中身は画像や Web フォントなど無害なものであった．

5.1.2　悪性データから得られた相関のある組み合わせ

悪性では UserAgent の変化と exe ファイルや jar ファイル，pdf ファイルを示す特徴に相関（mc1，mc2, mc3，mc15，mc16）が見られた．これらのファイルは攻撃に使用される，または，攻撃の結果ダウンロードさせられるものであり，攻撃時やマルウェアのダウンロード時に際して UserAgent が変化するため相関が見られたと考えることができる．

良性データと同様に，js ファイルを示す特徴とホスト

表 2　相関のある組み合わせ
Table 2　Correlated Combination

ID	良性	ID	悪性
bc1	ホストの遷移，Web 階層 5 以上	mc1	UserAgent の変化，exe を示すマジックナンバー
bc2	拡張子.js，Web 階層 5 以上	mc2	UserAgent の変化，拡張子.jar
bc3	ホストの遷移，拡張子.js	mc3	UserAgent の変化，exe を示す MIME タイプ
bc4	X-Powered-By ヘッダの値，Web 階層 5 以上	mc4	ホストの遷移，拡張子.js
bc5	exe を示す MIME タイプ，拡張子.js	mc5	ホストの遷移，js を示す MIME タイプ
bc6	ホストの遷移，js を示す MIME タイプ	mc6	受信データ量，拡張子.js
bc7	ホストの遷移，exe を示す MIME タイプ	mc7	受信データ量，js を示す MIME タイプ
bc8	受信データ量，js を示す MIME タイプ	mc8	受信データ量，ホストの遷移
bc9	受信データ量，拡張子.js	mc9	X-Powered-By ヘッダの値，拡張子.jar
bc10	X-Powered-By ヘッダの値，ホストの遷移	mc10	拡張子.pdf，拡張子.jar
		mc11	拡張子.swf，拡張子.pdf
		mc12	拡張子.swf，拡張子.pdf
		mc13	swf を示す MIME タイプ，拡張子.pdf
		mc14	exe を示す MIME タイプ，pdf を示す MIME タイプ
		mc15	UserAgent の変化，pdf を示す MIME タイプ
		mc16	UserAgent の変化，jar を示す MIME タイプ
		mc17	X-Powered-By ヘッダの値，Web 階層 5 以上
		mc18	X-Powered-By ヘッダの値，pdf を示す MIME タイプ
		mc19	X-Powered-By ヘッダの値，UserAgent の変化

の遷移や受信データ量についても相関（mc4，mc5，mc6，mc7）が見られた．5.1.1 節で触れた内容に加えて，悪性データでは JavaScript を用いた攻撃サイトへのリダイレクトによるホストの遷移や，その後の攻撃，マルウェアのダウンロードで大きなファイルを受信するためだと考えれる．

pdf ファイルを示す特徴と swf ファイル，jar ファイル，exe ファイルを示す特徴でも相関（mc10，mc11，mc12，mc13，mc14）が見られた．悪性データを確認したところ，全 983 件中 278 件のデータで PDF ファイルが存在しており (良性では 765 件中 1 件)，攻撃時に pdf ファイルが利用されるため出現数が多く，相関も高くなったと考えれる．

6. アソシエーション分析

データ中の検知項目の出現の有無からアソシエーション分析を行う．アソシエーション分析とはデータ内のアイテムの関連性を表す相関ルールを抽出するデータマイニング手法である．相関ルールは X ⇒ Y のように表され，X を条件部，Y を結論部と呼ぶ．条件部 X，結論部 Y には単一のアイテムだけでなく，複数のアイテムを含むこともある．相関ルールは支持度 ($supp$)，信頼度 ($conf$)，リフト値 ($lift$) という評価指標を持ち，分析者はこれらの指標を指定することで目的とする相関ルールを抽出することができる．

支持度は，ルールが全体の中でどのくらい出現する割合が高いかを表し，全トランザクション N のうち，ルールの条件部 X と結論部 Y を同時に含む確率となる．

$$supp(X \Rightarrow Y) = \frac{X \cap Y}{N}$$

信頼度は，条件 X を含むトランザクションのうち，結論 Y も同時に含む確率となる．これはルールの関連性の強さを表し，信頼度が大きければ大きいほど条件 X が出現した際に結論 Y も出現する可能性が高いといえる．

$$conf(X \Rightarrow Y) = \frac{X \cap Y}{X}$$

リフト値は条件 X と結論 Y の信頼度を結論 Y の出現率で割った値である．リフト値が大きい相関ルールは条件 X と結論 Y には関連性があり，リフト値が小さい相関ルールは条件 X と結論 Y には関連性がうすく，結論 Y は条件 X によらないものであるといえる．

$$lift(X \Rightarrow Y) = \frac{conf(X \Rightarrow Y)}{Y}$$

ここでは相関ルール抽出アルゴリズムとして Apriori アルゴリズムを用いた．トランザクション中のアイテムの種類が多ければ，アイテムの組み合わせ数も増え，相関ルールの数は膨大になる．Apriori アルゴリズムでは最小支持度と最小信頼度を与えることで，それらを満たす相関ルールだけを効率的に抽出することができる．

6.1 アソシエーション分析のためのデータの加工

アソシエーション分析を行うために量的変数である検知項目の出現数を質的変数に変換する．1 つの Web サイトにアクセスした際のリダイレクトチェインについて，検知項目が出現したかどうかでバスケットデータを作成する．例えば，「Web 階層 1 の Web ページが html ファイルであり，Web 階層 2 で js ファイル，Web 階層 3 で別の js ファイルが出現し，Web 階層 2 から Web 階層 3 の間にホストの遷

表 3　良性データから得られた相関ルール
Table 3　Rules Extracted Form Benign Data

ID	supp	conf	lift	条件	結論
b1	0.13	0.99	1.17	Web 階層 5 以上	ホストの遷移
b2	0.12	1.0	1.12	Web 階層 5 以上	js ファイル
b3	0.70	0.97	1.08	受信データ量	js ファイル
b4	0.11	0.92	1.03	X-Powered-By ヘッダの値	js ファイル
b5	0.10	0.87	1.03	X-Powered-By ヘッダの値	ホストの遷移
b6	0.62	0.86	1.02	受信データ量	ホストの遷移
b7	0.77	0.91	1.01	ホストの遷移	js ファイル
b8	0.77	0.85	1.01	js ファイル	ホストの遷移

表 4　悪性データから得られた相関ルール
Table 4　Rules Extracted Form Malignant Data

ID	supp	conf	lift	条件	結論
m1	0.12	0.93	3.75	jar ファイル	UserAgent の遷移
m2	0.12	0.91	1.30	jar ファイル	X-Powered-By ヘッダの値
m3	0.22	0.87	1.23	UserAgent の変化	X-Powered-By ヘッダの値
m4	0.23	0.84	1.19	Web 階層 5 以上	X-Powered-By ヘッダの値

移が生じる」場合のバスケットデータの内容は「js ファイル，ホストの遷移」となる．作成したバスケットデータを用いてアソシエーション分析を行った．

6.2　抽出された相関ルール

悪性データ，良性データ，それぞれからアソシエーション分析を行い，支持度が 0.05 以上，信頼度が 0.8 以上，リフト値が 1.0 以上の相関ルールを抽出した．良性データから抽出したルールを良性ルール，悪性データから抽出したルールを悪性ルールとする．良性，悪性ルールのうち，Web 階層を条件部，結論部両方に持つるようなルール（例，Web 階層 3⇒Web 階層 2）は，因果関係が明らかであり，悪性，または，良性通信特有の特徴を表すものではないので除外した．また，良性，悪性ルール中の両方で見られたルールについても除外した．その結果，良性ルールは 25 個，悪性ルールは 178 個抽出された．良性ルール，悪性ルールについてその一部分を示す (**表 3**，**表 4**).

6.2.1　良性ルール

良性では，「Web 階層 5 以上 ⇒ ホストの遷移」，「Web 階層 5 以上 ⇒js ファイル」のルール（b1，b2）が抽出された．「js ファイル ⇒ ホストの遷移」，「ホストの遷移 ⇒js ファイル」（b7，b8）も見られる．b7，b8 は支持度が高く，ホストの遷移と js ファイルは高い出現率で出現する組み合わせであるといえる．5 章でも述べたように，近年の Web において JavaScript は活発に利用され，ホストの遷移も生じやすいため，b7，b8 はその影響によるものだと考えれる．

また，「X-Powered-By ヘッダの値 ⇒js ファイル」，「X-Powered-By ヘッダの値 ⇒ ホストの遷移」のルール（b4，b5）も見られた．データを確認したところ，良性サイトでも多様なサイトでヘッダ内に古い PHP のバージョン情報を含んだ HTTP 応答が存在していた．js ファイル，ホストの遷移は頻繁に出現するため，古い PHP のバージョン情報が出現した際に，同時に js ファイル，ホストの遷移が出現する割合が高く，b4，b5 のルールが抽出されたと考えれる．

6.2.2　悪性ルール

悪性ルールでは「jar ファイル ⇒UserAgent の変化」，「jar ファイル ⇒X-Powered-By ヘッダの値」，「UserAgent の変化 ⇒X-Powered-By ヘッダの値」のルール（m1，m2，m3）が見られた．Exploit Kit が Java の脆弱性を利用する場合，Java アプリケーションによる攻撃サイトへの通信が発生する．また，攻撃サイトでは X-Powered-By ヘッダの値に古い PHP バージョンの情報が含まれる場合がある．上にあげたルールは，これらの特徴を反映されたものだと考えられる．特に m1 ではリフト値が高く，強い関連性があると言える．

7.　おわりに

本研究では，Drive-by-Dwonload 攻撃検知において既存の検知手法が用いる攻撃時の通信データの特徴をまとめ，それらの関連性を分析した．相関分析では，良性・悪性データともに Web 階層 5 以上の Web ページの出現数とホストの遷移や js ファイルの出現数に相関があることが分かった．また，良性サイトから抽出した相関ルールにおいても，「Web 階層 5 以上 ⇒ ホストの遷移」，「受信データ量 ⇒js ファイル」などのルールが見られた．これは良性・悪性データに限らず頻繁にホストの遷移，js ファイルが出現するためだと考えられる．Drive-by-Dwonload 攻撃検知において，Web 階層が深いかどうかとホストの遷移や js ファイルの組み合わせでは，検知手法の精度向上に寄与していないと考えられる．

悪性データ中では UserAgent の変化と exe ファイル，jar ファイル，pdf ファイルといった攻撃に関連するファイルタイプの間に相関が見られた．また，悪性データを用いたアソシエーション分析おいて，「jar ファイル ⇒UserAgent の変化」といったルールが高いリフト値で得られており，Drive-by-Dwonload 攻撃検知において，UserAgent の変化，jar ファイルの組み合わせは有効であると言える．さらに，UserAgent の変化と攻撃に関連するファイルタイプの組み合わせも検知の際に有効に働くと考えられる．また，悪性データでは良性ではほとんど出現することのない pdf ファイルが出現し，その出現は exe ファイルや jar ファイルの

出現と相関があることが分かった．その結果，pdf ファイルとその他の攻撃に関連するファイルから攻撃の検知を行える可能性がある．

参考文献

[1] エクスプロイトキット最新動向分析：Web サイト改ざんと不正広告を経由し、Flash 脆弱性を攻撃, http://blog.trendmicro.co.jp/archives/13017.

[2] 北野美紗, 大谷尚通, 宮本久仁男ほか, Drive-by-Download 攻撃における通信の定性的特徴とその遷移を捉えた検知方式, 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.595–602, 2013.

[3] 酒井裕亮, 佐々木良一ほか, Drive By Download 攻撃に対する HTTP ヘッダ情報に基づく検知手法の提案, 研究報告コンピュータセキュリティ (CSEC), Vol.2013, No.29, pp.1–6, 2013.

[4] 安藤槙悟, 寺田真敏, 菊池浩明, 趙晋輝ほか. 通信の遷移に着目した不正リダイレクトの検出による悪性 Web サイト検知システムの提案. 研究報告コンピュータセキュリティ (CSEC), Vol.2011, No.32, pp.1–6, 2011.

[5] 佐藤祐磨, 中村嘉隆, 高橋修ほか, 通信遷移と URL の属性情報を用いた悪性リダイレクト防止手法, コンピュータセキュリティシンポジウム 2015 論文集, Vol.2015, No.3, pp.8–15, 2015.

[6] 工藤聖, トラン・コン・マン, 中村康弘ほか, HTTP リクエストシーケンスに注目した不正リダイレクトの検出, コンピュータセキュリティシンポジウム 2015 論文集, Vol.2015, No.3, pp.221–225, 2015.

[7] 寺田成吾, 小林峻, 小出和弘, 羽藤逸文, 瀬戸口武研, 道根慶治, 山下康一ほか, ネットワーク通信の相関性に基づく Drive-by Download 攻撃検知手法, コンピュータセキュリティシンポジウム 2015 論文集, Vol.2015, No.3, pp.1–7, 2015.

[8] 高田雄太, 寺田真敏, 村上純一, 笠間貴弘, 吉岡克成, 畑田充弘ほか, ”マルウェア対策のための研究用データセット MWS Datasets 2016”, 研究報告セキュリティ心理学とトラスト (SPT), Vol.2016, No.17, pp.1–8, 2016.

[9] Alexa Top 500 Global Sites, https://www.alexa.com/topsites.