Computer Security Symposium 2019
21 - 24 October 2019

ブラウジング履歴情報に基づく 悪性サイトの事前検知

巻島 和雄[1,a)]　三須 剛史[1]

概要: 情報収集等を目的とする Web ページへのアクセスは多くの人が日常的に行っていることであり，常に Web 媒介型の攻撃に晒される危険性がある．個人情報の窃取，マイニングマルウェアへの感染といった Web 上の悪性コンテンツへの対策としては危険なページへのアクセスをブロックするという方法が一般的であるが，事前に警告することで悪性コンテンツへのアクセスを防ぐことが出来るのであればより望ましい．本研究では，ブラウジング履歴を用いてユーザが悪性コンテンツへアクセスする前に検知を行う手法を提案した．実験では，WarpDrive 実証実験データから悪性コンテンツを含むサイトを閲覧した際のブラウジング履歴を抽出し，どのような傾向があるか調査を行った．加えて，同実証実験データにおいて悪性コンテンツを含まないブラウジング履歴を良性サンプルとして，履歴情報に基づく悪性コンテンツの事前検知が可能であるか実験を行った．実験の結果，対象ページに遷移する 1 ホップ前から 3 ホップ前の履歴より抽出した特徴量を用いることで，対象ページ自身の情報を用いずとも高い精度で悪性コンテンツの検知が可能であることを示した．

キーワード: 悪性サイト 検知，機械学習

Estimate of Web Content Maliciousness Using Browsing Histry Data

Kazuo Makishima[1,a)]　Takeshi Misu[1]

Abstract: Getting information by web-browsing is very common and daily basis on many people, and there is always a risk of being exposed to Web-threats. To prevent malicious content on the Web, such as theft of personal information and infection with mining malware, block access to such pages with blacklist is usual method. But it is more desirable if we can prevent access to malicious content by warning in advance. In this study, we proposed a method that estimate malicious content before user access. In experiment, by using WarpDrive demonstration experiment data, we examined tendency of browsing history data which contain access with malicious content. In addition, along with browsing history data which does not contain access with malicious content, we construct malicious content access classifier and verify that accuracy. Experients shows that even not using information from page that contain malicious content, only with browsing histry data which contain 1 to 3 hop before the malicious page, we can estimate malicious content at high accuracy.

Keywords: Malicious website detection, Machine learning

1. 研究の背景

Web ブラウザを経由してインターネット上のコンテンツにアクセスすることは一般的であるが，そこには多くの脅威が存在する．ドライブバイダウンロード攻撃のようにマルウェアをダウンロードさせる事により損害を及ぼすパターンだけでなく，ユーザ自身に個人情報を入力させることを促すフィッシング攻撃や，無許可で仮想通貨の採掘にマシンパワーを浪費させるマイニング攻撃など，その種類は幅広い．

こういった様々な脅威に対応する方法として，大別して二つのアプローチが考えられる．一つは悪性のコンテンツそのものに着目し，同じ内容のコンテンツや共通した特徴を持つコンテンツを検知しブロックする手法であり，もう

[1] 株式会社セキュアブレイン
SecureBrain Corporation
[a)] kazuo_makishima@securebrain.co.jp

ひとつは悪性コンテンツの IP アドレスやドメインに着目し，悪性コンテンツへの接続を遮断する手法である．

後者の手法についても既知の悪性コンテンツの存在する URL をブラックリストとして利用するだけでなく，既知の悪性コンテンツの存在する URL と共通した特徴を持つ未知の URL に対してその危険性を推測するような方法が提案されている [1].

特徴を抽出しての推測が可能な理由として，Web 上において悪性コンテンツの存在する場所には一定の濃淡と傾向が存在するという事実がある [2]．ユーザを誘引しやすいコンテンツを含むサイトであったり，コンテンツの健全性を維持する機構が充分に働いていないようなサイトにおいては悪性コンテンツに遭遇する蓋然性が高くなる．

このような偏りが存在することを利用すれば，悪性コンテンツの多く存在する場所への接近経路を監視することによって，ユーザが悪性コンテンツにアクセスすることを未然に防ぐことが可能であると考えられる．

我々の研究では，ブラウジング履歴情報を用いてユーザが悪性コンテンツへアクセスする経路の調査を行った．加えて，悪性コンテンツへアクセスする前の段階においてその危険性を検知することが可能であるか検証するため，悪性コンテンツを検知する分類器を作成しその精度を評価した．

本論文の構成を以下に示す．第 2 章は関連する研究について述べる．第 3 章では本研究で利用したデータセットについて，その収集方法の説明と内容の分析を行う．第 4 章においては前章の内容を踏まえつつ悪性コンテンツ判別のための分類器を生成する実験を行う．第 5 章にて実験結果について述べ，第 6 章でその内容を踏まえ今後の展望を述べる．

2. 先行研究

URL の特徴量を利用した悪性サイト検知については，孫 [1] らが Bayesian Sets を利用した既知の悪性 URL と類似した特徴を持った URL の探索について報告している．佐藤 [3] らは，Drive-by-Download 攻撃に使われている ExploitKit 毎に利用される悪性 URL が異なる特徴を持つことに着目し，良性／悪性の判別に加え攻撃に使われた ExploitKit の判別も可能であることを示している．

ページ遷移に伴う URL 列を検知に活用した例としては，山西 [4] らによるリダイレクトチェインに含まれる URL 群を入力とした研究がある．

また，Web 上における脅威が偏りを持って分布していることについては斉藤 [2] が IPv4 空間を視覚的にプロットし，危険な領域を明らかにするインターフェースを提案している．

これら先行研究に対する本研究の独自性としては，遷移タイプの情報を検知に利用したことと，対象となるページの情報を用いずに予測的な検知を行う場合を想定したことが挙げられる．

3. データセット

悪性コンテンツを含むサイトにアクセスした際のブラウジング履歴情報を取得するため，WarpDrive 実証実験のデータを利用した．

WarpDrive（ Web 媒介型攻撃対策技術の実用化に向けた研究開発）は，Web 媒介型攻撃の実態把握と対策技術の向上のための研究開発プロジェクト [5] である．

実証実験は 2018 年 6 月より開始しており，Chrome 拡張機能の形で広く一般ユーザに向けタチコマ・セキュリティ・エージェントの配布を行っている．これはブラウザに常駐し危険なサイトへの接続を防ぐと共にユーザの同意を得てブラウジング履歴情報の収集を行うものである．

2019 年 7 月の時点で累計の登録ユーザ数は 8000 を超え，日毎 800 程度のユニークユーザ ID から 1000 万規模のブラウジング履歴情報を収集している．

3.1 データ収集基準

本研究においては収集された WarpDrive 実証実験データのうち 2019 年 2 月から 5 月まで 4 ヶ月分のデータを利用し，内容の分析と悪性コンテンツ検知実験を行った．

悪性データとして，4 ヶ月分の WarpDrive 実証実験データから悪性コンテンツを含むサイトへの訪問履歴を全て収集した．

悪性コンテンツを含むサイトの定義としては，ページに含まれるコンテンツのうち一つでも Google Safe Browsing[6] の検知で悪性と判定されたものとした．

ブラウジング履歴の収集範囲については，有効なページ遷移数の確保とデータ取得に要する時間コストを考慮し，悪性コンテンツを含むサイトにアクセスした時点から遡って 20 分前までのページ遷移を追跡するものとした．

WarpDrive 実証実験データでは，ページの遷移が行われるに際してどのような要因で遷移が行われたかが記録される．記録されるタグとその意味については**表 1** に示す．

このうち，遷移タイプが “generated”，“auto_bookmark”，“start_page”，“typed” で表されるものについてはページ内の情報とは関連しない外部の情報による遷移である．よって，これらの遷移タイプが記録されていた場合それ以前のブラウジング履歴は利用しないものとした．

検知実験においては対照データとして悪性コンテンツとの接触を含まないブラウジング履歴情報を収集する必要がある．これについては，母集団の差異によるデータの偏りを防止するため，悪性データと同じく 2019 年 2 月から 5 月までのデータのうち以下の条件をすべて満たすものを対象とした．

- 対象期間である 4 ヶ月のうちに一度でも悪性コンテン

表 1　遷移タイプ一覧

値	説明
auto_bookmark	ブックマークまたはアイテムをクリックした
form_submit	フォームを送信した
generated	アドレスバー入力から出現した候補をクリックした
link	別のページのリンクをクリックした
reload	リロードした・セッションが復元された
start_page	コマンドラインで指定されたか，または開始ページ
typed	URL をアドレスバーに入力した
・以下の値は修飾句であり 他の値と共に付与される場合がある	
client_redirect	ページ上の JavaScript またはメタリフレッシュタグに起因する 1 つまたは複数のリダイレクトが発生
server_redirect	サーバーから送信された 3XX HTTP ステータスコードに起因する 1 つまたは複数のリダイレクトが発生
forward_back	[進む] または [戻る] ボタンを使用してナビゲーションを開始
from_address_bar	アドレスバーからナビゲーションを開始

ツを含むサイトにアクセスしたユーザ

- 悪性コンテンツを含むサイトにアクセスしていない日のブラウジング履歴

上記の条件を満たすデータからランダムにページへの訪問情報を抽出し，対象となったページを基点として 20 分前までの履歴情報を収集することで良性データとした．

以上の基準により悪性 3585 件，良性 3301 件のブラウジング履歴情報を抽出した．

3.2　悪性コンテンツを含むサイトへのアクセスにおける傾向分析

悪性コンテンツを含むサイトへのアクセスが行われる際の傾向を把握し検知実験の方針を定めるため，抽出したブラウジング履歴情報を分析した．

ユーザ毎のアクセス頻度

悪性コンテンツを含むサイトへのアクセス頻度をユーザ単位で集計すると，4 ヶ月間の観測で確認されたユニークユーザ ID は 263 であった．実証実験全体でのアクティブユーザ数から類推すると，期間中にブラウジングを行ったユーザの 2 割程度が少なくとも一度，悪性コンテンツを含むサイトにアクセスを行っていることになる．

ユーザ毎に，期間中に悪性コンテンツを含むサイトに何回アクセスしたかヒストグラムとして集計した結果を**図 1** に示す．縦軸がカラム毎のユーザ数，横軸が対象ユーザが期間中に悪性コンテンツを含むサイトにアクセスした回数を示す．

悪性コンテンツを含むサイトに 1 回しかアクセスしていないユーザが 100 名で全体の 40 %程度を占めるのに対し，少数ではあるが継続的に悪性コンテンツを含むページへのアクセスを行っているユーザも存在した．

時間的情報

一日のうちどの時間帯でのアクセスが多いかについて，悪性コンテンツを含むサイトを**図 2**，良性のサイトを**図 3** に示す．縦軸が一時間単位で区切ったデータ出現数であり，横軸は日本時間である．

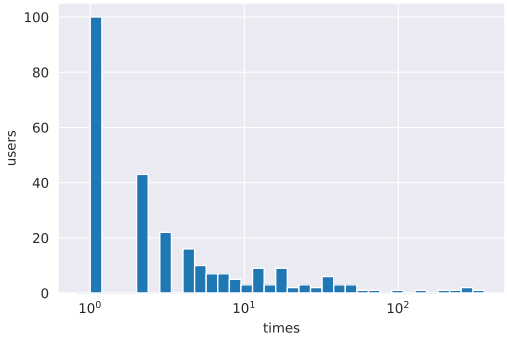

図 1　ユーザ毎アクセス回数

今回利用したデータセットは WarpDrive 実証実験参加者を対象としたものであり，その性質上私用のデバイスでのブラウジング履歴情報が多くなるものと考えられる．夕方から夜にかけてアクセス数が多くなっている傾向はその反映として自然なものといえる．

収集対象となったユーザ集団が同一であるにもかかわらず，図 2 と図 3 の間では差異が存在する．午後から夕方の時間帯においては悪性コンテンツを含むサイトへのアクセス頻度が良性サイトのものに比してあまり伸びず，25 時頃を中心とする深夜帯において悪性コンテンツを含むサイトにアクセスする比率が上昇している．

頻出ドメイン

アクセスが観測された悪性コンテンツを含むサイトについて，ドメイン単位で出現頻度の上位を取ると**表 2** のようになる．傾向として，比較的長時間の滞在が想定される動画系のサイトが多く見られる．他には，第 5 位にプログラミング関係のサイトがある．このサイトは他のプログラミング関係の Q&A サイトの内容をコピーしたうえで低品質の機械翻訳に通したものを掲載しているサイトであり，関連語句をキーとして検索したユーザを誘引する意図があると思われるが，どういった目的によって運営されているかは不明である．

幾つかのサイトについてアクセスが検出された回数が時

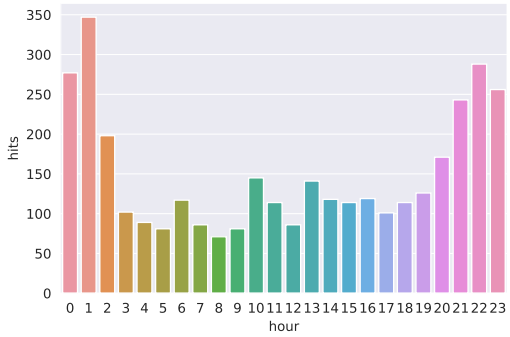

図 2　アクセス時刻　悪性コンテンツを含むサイト

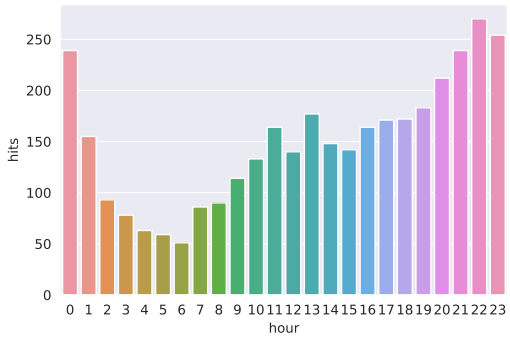

図 3　アクセス時刻　良性のサイト

期によってどのように変化しているかを図 4，図 5 に示す．横軸は調査対象期間の 2019 年 2 月から 5 月までを 10 日刻みで分割したものであり，縦軸は期間中何回悪性コンテンツを含むサイトとして出現しているかである．

＃漫画サイトについては 2019 年 4 月頃にサイトが閉鎖されたという情報があり，悪性コンテンツの検知回数も以降は 0 となっている．

＃プログラミング質問サイト 翻訳については対象調査期間の途中から悪性コンテンツを含んでいるとして検知されるケースが観測されるようになっている．しかし，InternetArchive[7] の情報によるとサイト自体は以前より存在していた．これは，新たに悪性コンテンツが配置されたか，Google Safe Browsing の検知対象が広がったなどの原因が考えられる．

このように，悪性コンテンツが検出されるサイトについては時期によってその出現率に変化が見られる．

アクセス経路

出現頻度の高いサイトについて，どのような経路を辿ってアクセスされているかを表 3，表 4 に示す．

表 3 は対象ページの 1 ホップ前がどのようなものであったかを示す．

“直接遷移” は表 1 で述べたページ内情報に関連しない遷移である．ブックマークからの遷移や URL をアドレスバーに入力しての遷移が含まれる．

“時間超過” は収集したブラウジング履歴情報に前ページの情報がなかったものである．収集範囲である 20 分以上

表 2　頻出ドメイン名上位 10

悪性サイト	回数	ユーザ数
＃動画サイト A	1983	59
＃漫画サイト	427	36
＃動画サイト B	253	8
＃動画サイト C	146	8
＃プログラミング質問サイト 翻訳	114	41
＃動画サイト D	56	6
＃動画サイト E	54	4
＃動画サイト F	29	1
＃ファイル共有サイト	25	4
＃動画サイト G	22	1

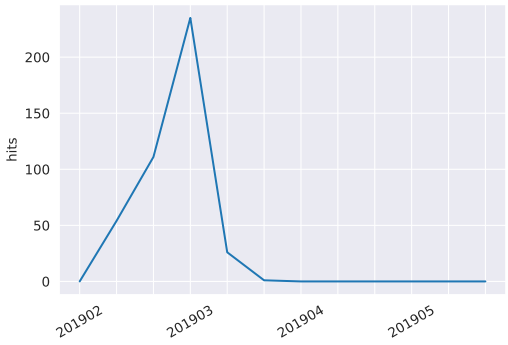

図 4　期間毎アクセス数　漫画サイト

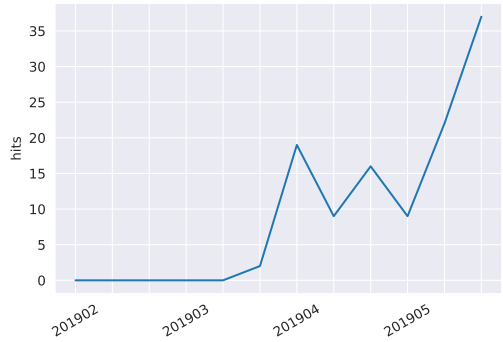

図 5　期間毎アクセス数　プログラミングサイト

の間当該ページに滞在していたケースであると考えられる．

“同一ドメイン”，“別ドメイン” は前ホップのページ URL が取得できたものであり，それぞれ対象の悪性コンテンツを含むページと同一のドメイン内のページから遷移してきたもの，異なるドメインから遷移してきたものを示す．

この結果を見ると，＃動画サイト A や＃動画サイト E 等で悪性コンテンツの存在するページ内に長時間滞在するケースの多いことがわかる．また，直接遷移の比率について＃ファイル共有サイトの値が際立って高いものとなっている．これは，ファイル共有ソフトなどの外部ツールを使って必要としているファイルの存在する URL を検索し，確認のためにブラウザからアクセスするといったような遷移経路が考えられる．

どのサイトでも同一ドメイン内から遷移してきた場合が多い．他ドメインからの遷移について分析するため，1

表 3　1 ホップ前からの遷移情報

悪性サイト	総観測数	直接遷移	時間超過	同一ドメイン	別ドメイン
# 動画サイト A	1983	98	758	957	170
# 漫画サイト	427	98	118	188	23
# 動画サイト B	253	25	74	96	58
# 動画サイト C	146	5	16	120	5
# プログラミング質問サイト 翻訳	114	3	28	76	7
# 動画サイト D	56	11	3	27	15
# 動画サイト E	54	1	32	11	10
# 動画サイト F	29	4	15	10	0
# ファイル共有サイト	25	18	1	6	0
# 動画サイト G	22	0	3	19	0

表 4　別ドメインからの遷移情報

悪性サイト	総観測数	直接遷移	時間超過	検索サイト	その他
# 動画サイト A	1983	517	1157	180	129
# 漫画サイト	427	172	175	17	63
# 動画サイト B	253	40	139	36	38
# 動画サイト C	146	21	77	0	48
# プログラミング質問サイト 翻訳	114	3	28	82	1
# 動画サイト D	56	12	3	3	38
# 動画サイト E	54	8	32	0	14
# 動画サイト F	29	13	16	0	0
# ファイル共有サイト	25	24	1	0	0
# 動画サイト G	22	0	13	9	0

ホップ前に限定せずブラウジング履歴情報を分析した結果を表 4 に示す．

表 3 が対象となる悪性コンテンツが存在するページの 1 ホップ前のデータであったのに対し，表 4 は対象ページの存在するドメインを単位としてそのドメインにどうやって入ってきたのかを分析した結果である．

表 4 おいて “直接遷移” は履歴を遡った結果対象ページと同ドメイン内の何処かのページに直接的に遷移していたことを示す．

同様に，“時間超過” は 20 分間以上当該ドメイン内でのブラウジングが継続されていたことを示す．

“直接遷移”，“時間超過” のどちらにも該当しない場合，別のドメインから遷移してきたことになる．これらについては，遷移元の URL から “検索サイト ” と “その他” に分けて記載した．

表 4 より，# プログラミング質問サイト 翻訳のように検索サイトからの遷移が大半であるサイトがある一方，# 動画サイト C や # 動画サイト E のように検索サイトからの遷移が皆無のサイトもある．

“その他” に含まれるドメインの内容としては，# 動画サイト C では動画ファイル配信に利用している CDN サービスのアドレスが，# 漫画サイトにおいては同じようにコミック類がアップロードされているサイトのアドレスが見られる等，サイトによってそれぞれ特徴がみられる．

複数の動画サイトに共通して現れる遷移元も存在し，動画のダウンロードツールを提供しているサイトやポータル的な役割をしているサイトが該当する．こういったサイトは自身が悪性のコンテンツを含まない場合でも様々な悪性コンテンツを含むページに遷移する可能性があり，検知における指標となる可能性がある．

4. 実験

ブラウジング履歴情報から対象のページが悪性コンテンツを含んでいるか否か検知することが可能であるかを検証するために機械学習の手法を用いて分類器の生成を行った．

ブラウジング履歴情報からの特徴量抽出を行うため，検知実験の対象データはある程度の長さの履歴情報を追うことのできたデータに限定する．今回の実験では 3 ホップ分の履歴を遡ることができたデータを対象とした．

この結果，悪性データ 3585 件のうち 1114 件，良性データ 3301 件のうち 1239 件で 3 ホップ分の履歴情報を追うことが出来たため，これを検知実験データセットとした．

4.1 実験条件

実験条件としては，対象となるページの情報を用いずに悪性コンテンツを含むページを検知できるかとブラウジング履歴情報の活用による検知精度の変化を検証できるよう，下記の 4 つの条件で実験を行った．

- 対象となるページの情報のみを使う場合
- 対象となるページの情報に加え，1 ホップ前から 3 ホッ

プ前のページの情報を使う場合

- 対象となるページの情報を使わずに，1 ホップ前のページの情報のみ使う場合
- 対象となるページの情報を使わずに，1 ホップ前から 3 ホップ前のページの情報を使う場合

なお，記述の簡略化のため以降上記の各実験条件について対象ページを 0 ホップ前と置き，利用したホップ数からそれぞれ [0]，[0, 1, 2, 3]，[1]，[1, 2, 3] と表記する．

4.2 利用した特徴量

ブラウジング履歴情報における各ページごとに下記の特徴量を取得した．

URL に含まれる特徴

汎用性の高い学習が行われることを期待し，特定の単語等を利用するのではなく文字列の長さや種別ごとの文字数等抽象化された特徴量を取得した．

・ URL 文字列の長さ

URL をドメイン部分/パス部分/クエリ文字列の 3 つに分割し，各部分の文字数+URL 全体の文字数で 4 特徴量とする．

・ URL 文字列に含まれる数字の数

URL をドメイン+パス部分/クエリ文字列部分の 2 つに分割し，各部分に含まれる半角数字 [0-9] の個数で 2 特徴量とする．

・ URL 文字列のトークンへの分割

URL 文字列を特定の区切り文字で分割することによりトークンを生成する．今回の実験においては区切り文字として [“&”, “%”, “/”, “?”, “=”, “-”, “_”, “.”] を採用した．

URL をドメイン+パス部分/クエリ文字列部分の 2 つに分割し，各部分をトークンに分割する．それぞれに対し，

- トークンの個数
- トークンごとの文字数の平均値
- トークンごとの文字数の最大値

を特徴量とする．

トークンを利用した特徴量は合計 6 特徴量となる．

URL の処理について例示する．`http://example.com/number_123/somewhat?q=css2019&d=url` を与えられた場合，

ドメイン部分　`example.com`

パス部分　`/number_123/somewhat`

クエリ文字列　`q=css2019&d=url`

に分割される．クエリ文字列をトークンに分割すると [q, css2019, d, url] となる．

遷移方式による特徴

表 1 の遷移タイプのうち，どのタイプの遷移であったかを One-Hot ベクトル化し特徴量とした．修飾句の出現パターンを網羅した結果，今回のデータセットにおいては 15 特徴量となった

ページ滞在時間

http_request の発生した時刻について 0 ホップ目のページとの差を秒数で取ったものを特徴量とした．

ドメイン遷移の有無

対象ページの URL とその 1 ホップ前のページ URL のドメイン部分が一致しているかどうかについて，一致していれば 1，一致していない場合 0 とした．この特徴量は遷移前のページと遷移後のページ両方の情報を使える場合のみ利用した．

以上，ページあたり 29 個の特徴量を用いた．

なお，各特徴量はページごとに計算されるため，例えば 4 ページ分の特徴を用いる条件 [0, 1, 2 , 3] の場合，ドメイン遷移情報を除く 28 特徴量が 4 ページ分と [0-1] 間，[1-2] 間，[2-3] 間のドメイン遷移情報で合計 115 次元の特徴量を利用して対象となるページが悪性コンテンツを含んでいるか否かの判定を行うこととなる．

4.3 分類器

分類器としてはランダムフォレスト [8] を用いた．これは決定木のアンサンブル学習によって分類を行うアルゴリズムであり，計算量が比較的軽いことと，分類にあたって重視された特徴量についての見通しが良く結果の分析が容易なため採用した．

本実験では実装として Python 上で動作する scikit-learn[9] 内の RandomForestClassifier を利用した．

4.4 評価基準

分類器の性能を評価する基準としては以下の 4 つを用いる．

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Fmeasure = \frac{2Recall * Precision}{Recall + Precision} \tag{4}$$

数式中の略称について，TP(True Positive) は悪性データを悪性と正しく判定したもの，TN(True Negative) は良性データを良性と正しく判断したものの個数を示す．

同様に，FP(False Positive) は良性データを悪性と誤って判定してしまったもの，FN(False Negative) は悪性データを良性と誤って判定してしまったものの個数を示す．

評価指標のうち，Accuracy は全てのテスト対象データのうち正答したものの割合を示す．

式から見て取れる通り，Precision の高さは FP の少なさ，つまりは誤検知の少ないことを示す．同様に Recall の

表 5　RandomForest パラメタ

パラメタ名	説明	走査範囲
n_estimators	生成する決定木の数	10, 20, ... , 100
criterion	分割純度の算出方法	“gini” or “entropy”
max_depth	決定木の深さの限界値	None(Inf), 1, 2, 4, 6, 8, 10
min_samples_split	最小の分割可能ノードサイズ	2, 4, 10, 12, 16

表 6　実験条件毎の最適パラメタ

実験条件	n_estimators	criterion	max_depth	min_samples_split
0	20	“entropy”	None	4
0, 1, 2, 3	90	“gini”	None	2
1	80	“gini”	None	4
1, 2, 3	70	“entropy”	None	2

高さは FN の少なさ，検知漏れの少ないことを示す指標となる．F-measure は両者の調和平均であり，分類器の総合的な性能を示す．

4.5　パラメタの設定

各特徴量セットについて実験条件毎にグリッドサーチにより適したパラメタを設定した．サーチを行った RandomForest のパラメタとその範囲を**表 5** に，実験条件毎に得られた最良のパラメタセットを**表 6** に示す．

4.6　交差検定

分類器の性能を検証するためデータセットを 10 分割して交差検定を行った．分割内容による偏りを防ぐため交差検定全体を 5 回反復して行い，その平均を結果とした．

5.　結果

実験の結果得られた評価値を**表 7** に示す．

対象となるページの情報を使った実験条件 [0]，[0, 1, 2, 3] の場合正答率 94 %程度で，使わなかった実験条件 [1]，[1, 2, 3] の場合でも正答率 90 %程度で悪性コンテンツを含むページの検知が可能であることがわかった．

対象ページの情報を使わない条件での，つまりは「次にアクセスするページは危険なものであるか否か」を予め判定するという条件において 9 割程度の正答率が出ていることは有用な結果であると考えられる．

対して，条件 [0] と [0, 1, 2, 3]，条件 [1] と [1, 2, 3] の間では検知精度に有意な差は認められなかった．より多いホップ数のブラウジング履歴情報を用いることで検知精度の向上が見られるのではないかと予想していたが，今回構築した分類器においては利用するホップ数を増やしても精度が向上しないという結果になった．

対象ページの情報を使わない場合どのようなページの検知が難しくなっているかを調べるため，条件 [0, 1, 2, 3] と条件 [1, 2, 3] を対象として 5 回の反復の全てにおいて検知に失敗したケースを抽出し，比較を行った．

表 7　検知実験結果

実験条件	Accuracy	Precision	Recall	F-Measure
0	0.940	0.961	0.910	0.935
0, 1, 2, 3	0.938	0.961	0.905	0.932
1	0.904	0.931	0.860	0.894
1, 2, 3	0.902	0.930	0.857	0.892

条件 [0, 1, 2, 3] でのみ正しく検知し，条件 [1, 2, 3] で検知に失敗しているケースは 69 件存在した．悪性のコンテンツを含むサイトを見逃したパターンである False Negative は 51 件あり，その中には# プログラミング質問サイト 翻訳の URL が多く含まれていた．3 章で述べたようにこのサイトへのアクセスは多くが検索サイト経由であり，1 ホップ前は検索エンジンのページであることが多い．検索エンジンのページは良性側データ中の 1 ホップ前にも多く含まれており，特徴に共通する部分が多いため検知が難しいものと考えられる．

逆に条件 [1, 2, 3] でのみ正しく検知し，条件 [0, 1, 2, 3] で検知に失敗しているケースも 46 件存在した．こちらについてはあまり明確な傾向は見られなかったものの，表 2 に出現していない比較的低頻度で出現しているサイトについて False Negative となる場合が多い．

生成された分類器において，検知の際にどの特徴量が重視されたかを**表 8** に示す．なお，各項目の特徴量名に続く数値は対象ページから数えたホップ数を示す．

今回のデータセットにおける悪性サイトは# 動画サイト A の URL が多く含まれており，どの実験条件においてもドメイン長が重要な検知基準として扱われている．

また，精度としては殆ど差がなかったものの実験条件 [0, 1, 2, 3] や実験条件 [1, 2, 3] では 2 ホップ前の情報も重要な特徴量として使われている．

URL に関連しない情報としてはページ滞在時間が重視されている．また，遷移タイプ情報については上位の特徴量としては現れなかったものの，リダイレクトの有無等が検知基準として利用されていた．

表 8　検知の際に重視された特徴量上位 5 件

実験条件 [0]	実験条件 [0, 1, 2, 3]	実験条件 [1]	実験条件 [1, 2, 3]
ドメイン長 0	ドメイン長 0	ドメイン長 1	ドメイン長 1
ドメイン+パス部の数字数 0	ドメイン長 1	URL 全長 1	ドメイン長 2
パス長 0	ドメイン長 2	ページ滞在時間 1	URL 全長 1
URL 全長 0	ドメイン+パス部の数字数 0	パス長 1	ページ滞在時間 1
パス部平均トークン長 0	パス長 0	パス部平均トークン長 1	ページ滞在時間 2

6. 今後の展望

本研究で行った実験によってページ URL から抽出した情報とページ遷移についての情報を元にして悪性コンテンツを含むサイトの存在を事前に高精度で検知できることを明らかにした．

しかし，今回の手法は履歴情報を用いる都合上一定長の履歴を持つデータのみを対象として扱っており．実際に収集データの半数以上を検知実験データセット作成の際に除外している．

検知実験データセットの対象範囲外となったデータは

(1) ページでの滞在時間が長すぎるため履歴を追えなかったもの

(2) ブックマーク等から直接アクセスしているもの

があるが,(1) については今回対象ページ訪問時刻の 20 分前までと設定したブラウジング履歴取得時間長を伸ばすことによってある程度は対応可能である.(2) については今回の手法を適用できないが，ユーザ単位で考えた場合ブックマークを利用するのは 2 回目以降の訪問であり，初回の訪問は何処か別のサイトから遷移してきている可能性が高い．初回の訪問となる未知の悪性コンテンツを含むサイトに対処できるのであれば充分な有用性があると考える．

他にも，1 ホップ前が検索エンジンのページであるような場合において検知精度が落ちることが挙げられる．こういった場合に利用可能な特徴，例えば検索語と危険コンテンツへのアクセスとの関係などについてを検知に取り入れることが出来れば精度の向上が期待できる．これは今後の課題としたい．

今回の手法を実際にユーザへの警告として活用する場合，事前警告という性質上その警告内容はある程度漠然としたものになってしまう．今後の研究においては検知精度の向上と共に，警告に対してユーザがどのように反応するのか WarpDrive 実証実験環境を利用し実例を収集するなど，実用化に向けた改善を試みたい．

謝辞　本研究は，国立研究開発法人情報通信研究機構の委託研究「Web 媒介型攻撃対策技術の実用化に向けた研究開発」の成果の一部です．ご協力いただいた皆様に，深く感謝します．

参考文献

[1] 孫 博, et.al: 既知の悪性 URL 群と類似した特徴を持つ URL の検索, Computer Security Symposium 2014

[2] 斉藤 典明: インターネット空間の汚れ具合を観察するインタフェースの提案, 情報処理学会研究報告 Vol.2013-DPS-156 No.29

[3] 佐藤 祐磨, 中村 嘉隆, 高橋 修: エクスプロイトキットで利用される文字列特徴を用いた悪性 URL 検出手法の提案, 情報処理学会研究報告　Vol.2016-CSEC-72 No.25.

[4] 山西 宏平, et.al: 畳み込みニューラルネットワークを用いた URL 系列に基づくドライブバイダウンロード攻撃検知, Computer Security Symposium 2016

[5] WarpDrive　`https://warpdrive-project.jp/index.html`

[6] Google Safe Browsing `https://safebrowsing.google.com/`

[7] The Internet Archive `https://archive.org/`

[8] Breiman, Leo (2001). “Random Forests”. Machine Learning 45 (1): pp532

[9] scikit-learn `https://scikit-learn.org/stable/`