社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－ＤＰＳ－138 (39)
2009－CSEC－ 44 (39)
2009／3／6

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

田村 佑輔† 甲斐 俊文†† 佐々木 良一†

†東京電機大学
〒101-8457 東京都千代田区神田錦町 2-2
tamura@isl.im.dendai.ac.jp, sasaki@im.dendai.ac.jp

††パナソニック電工株式会社
〒108-0014 東京都港区芝 4-8-2 松下電工田町ビル８階
kai.toshifumi@panasonic-denko.co.jp

あらまし 近年, SQL インジェクションを用いて Web サイトに不正スクリプトを埋め込む改ざん攻撃が急増している．この攻撃は，サイトを閲覧したユーザにマルウェアを感染させることを目的としており，感染源が正規サイトであることから一般ユーザ側での対策が困難となっている．

本研究では，実際の改ざんサイトや不正スクリプトの調査を通して，サイトタイトルやスクリプトの記述パターンを分析することで，未知不正スクリプトを自動的に検出可能な方法を発見した．この方法を用いて，改ざんサイト及び不正スクリプトを検知するためのシステムを提案する．

Proposal of detection method using search engine against manipulation attack targeted at common users

Yusuke TAMURA† Toshifumi KAI†† Ryoichi SASAKI†

†Tokyo Denki University
2-2, Kanda-Nishiki-cho, Chiyoda-ku, Tokyo, 101-8457 JAPAN
tamura@isl.im.dendai.ac.jp, sasaki@im.dendai.ac.jp

††Panasonic Electric Works Co.,Ltd
4-8-2, Shiba, Minato-ku, Tokyo 108-0014, Japan
kai.toshifumi@panasonic-denko.co.jp

Abstract Recently, the manipulation attack to website embedding malicious script using SQL injection vulnerability is increasing. Purpose of this attack is to infect users PC which browse the site with the malware. It is difficult for users to protect this attack because source of infection is website that gets high evaluation. In this paper, we discovered automatic process of detect unknown malicious script, based on characteristics from an investigation. Moreover, we propose a method to detect in manipulation website and malicious script using the process.

1. はじめに

2000 年に発生した中央省庁の Web サイト改ざんに代表されるように，Web サイトへの改ざん攻撃は以前から行われていた．これらの改ざん攻撃は，サイトデザインを改ざんすることで，攻撃者自身の主張・メッセージを訴えることが目的であったといえる．しかし近年，Web サイトを閲覧した一般ユーザの PC にマルウェアを感染させることを目的として，サイトに不正なスクリプトを挿入する新たな Web サイト改ざん攻撃が増加している．2008 年 3 月には，日本をターゲットとした大規模な攻撃が行われ，多数の正規サイトが改ざんの被害に遭った．現在でも攻撃は継続しており，個人 HP や地方自治体などの多くのサイトが被害を受けている[1].

この改ざん攻撃に対し，サイト管理側での対策はもちろんのこと，標的となっている一般ユーザ側でも対策が求められている．しかし，下記のような理由からユーザ側での対策が困難になっている．

- マルウェア感染源が正規サイトである
- サイトを閲覧しただけでマルウェアに感染する可能性がある
- 視覚的な情報では改ざんの認識が難しい

本研究では，実際の改ざんサイト及び挿入された不正スクリプトを調査・分析し，そこから得られた特徴を用いて改ざんサイト・不正スクリプトを判別するための検知手法を提案する．なお本稿では，2 章でユーザ標的型改ざん攻撃について，3 章で既存の対策手法について，4 章で実施した調査について述べる．調査から得られた特徴を用いて，5 章で検知手法を提案し，6 章で考察を行い，最後に 7 章でまとめる．

2. ユーザ標的型 Web 改ざんについて

2.1 改ざん方法

ユーザ標的型 Web 改ざん攻撃における Web ページの改ざんは，SQL インジェクションと呼ばれる手法を用いて，無差別かつ広範囲に行われている．

SQL インジェクションとは，正規サイト上でデータベースと連携して運用されている Web アプリケーションの脆弱性を突き，データベースを不正に操作することで，データベース内の情報の不正取得や改ざんなどを行う行為である．ユーザ標的型改ざん攻撃では，この手法を用いて JavaScript や iframe などのスクリプトが不正に Web ページ内に挿入されている．

2.2 攻撃の流れ

Web サイトの改ざんから一般ユーザにマルウェアが感染するまでの流れを図 1 に示す．

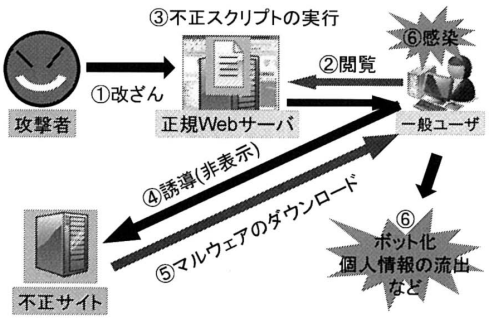

図 1　ユーザ標的型 Web 改ざんの流れ

まず，2.1 で述べた手法で正規サイトが改ざんされる(図 1,①)．このとき挿入されるのは，以下の図 2 のようなスクリプトである．

```
<script src=http://誘導先 URL><script>
<iframe src=http://誘導先 URL><iframe>
```

図 2：挿入される不正スクリプトの例

一般ユーザが改ざんされた正規サイトを閲覧した場合(図 1,②)，不正スクリプトが実行され(図 1,③)，スクリプト中に記述された不正サイトの URL 先に誘導・接続される(図 1,④)．不正スクリプトには誘導先のサイトが表示されないよう細工がされているため，一般ユーザは不正サイトに接続されていることを認識できない．そして，不正サイトからマルウェアがダウンロードされ(図 1,⑤)，一般ユーザの PC に脆弱性があれば感染してしまう(図 1,⑥)．このとき InternetExplorer, FlashPlayer など多種多様なアプリケーションの脆弱性が利用されるので，一部のアプリケーションのアップデートだけではこの攻撃に対処しきれない．

2.3 ユーザ側における対策の必要性

Web改ざん攻撃が発生する根本的原因は，Webサイト側の脆弱性にあるといえる．すなわち，Webサーバ管理者側でWebサイトの脆弱性をなくしてしまえば，改ざん自体を防ぐことが可能である．しかし，全てのWebサーバ管理者がこうした対策を行っているとは限らないため，ユーザ側でも対策をとる必要がある．Webにおける簡便なマルウェア対策として「怪しいページにアクセスしない」，「怪しいファイルをダウンロードしない」という人の手による対策方法が挙げられる．しかし，ユーザ標的型Web改ざん攻撃に対しては，下記のような理由から全く効果がなくなってしまう．

- 改ざんされた正規ページを閲覧しただけで感染の恐れがある
- 改ざんを直感的に認識できない
- 不正サイトへの誘導が不可視である

このようなことから，ユーザ側においては「機械的に改ざんサイト及び不正サイトへのアクセスを防止する」という対策が必要となってくる．

3. ユーザ側における既存対策手法

3.1 既存手法の概要

「改ざんサイト及び不正サイトへのアクセスの防止」という対策の具体的方法として，危険なサイトに対して事前に警告を出すGoogleのセーフブラウジング機能の活用や，不正サイトや不正スクリプトのブラックリストによるアクセス制限などが挙げられる．

3.2 問題点

セーフブラウジング機能の警告は，Googleのクローラが巡回した際にサイトの危険性を識別して出される．このことから，改ざんが行われてからクローラが巡回するまでの間は警告を出すことができないという問題点がある．ブラックリストによるアクセス制限は，この問題に対処することができるが，リストに掲載されていない未知の不正スクリプトに対しては対処できないという新たな問題が挙がる．

そこで我々は，この「未知の不正スクリプト」に対処するための方法として，不正スクリプト共通の特徴を用いて判定を行うことを考えた．次章では，その特徴を抽出するために行った調査・分析について述べる．

4. 調査

4.1 調査概要

本研究では，改ざんサイトを調査対象とした1次調査と，不正スクリプトを調査対象とした2次調査の2つを実施した．1次，2次共に，検索エンジンを用いて行い，特定の検索キーワードによる検索結果から改ざんサイトや不正スクリプトの収集し，分析を行った．

調査作業は，表1に示す環境下で自作のプログラムを用いて行った．検索エンジンとしてGoogleを採用した理由は，他のWeb検索エンジンに比べて改ざんサイトの検索ヒット数が多かったためである．なお，"Google AJAX Search API"とは，Googleの検索エンジンをプログラム上で使用するためのAPIのことである．

表1　調査環境および開発環境

OS	Windows XP Professional
CPU	Intel® Pentium® M processor 1100MHz
メモリ	760MB
開発環境	Visual C# 2005 Express Edition
	Google AJAX Search API
検索エンジン	Google

4.2 1次調査 (改ざんサイト調査)

4.2.1 調査目的および方法

1次調査の目的は，改ざんサイト及び不正スクリプトにおいての共通の特徴を発見することである．

改ざんサイトの検索で用いられる検索キーワードは，スクリプトの基本書式と，Shadowsever Foundationにて公開されている不正誘導先ドメインのブラックリスト[2]を組み合わせたものである．調査では，この検索キーワードを用いてサイト検索を行い，検索結果から改ざんページを収集し，特徴の分析を行った(図3).

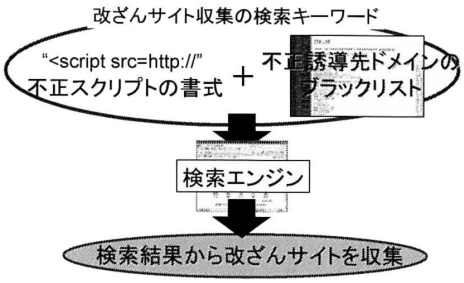

図 3　改ざんサイト収集方法

4.2.2　調査結果

1 次調査では 20,439 件の改ざんページを収集することができた．また，サイト内容を確認したところ，個人 HP のほか，企業，公益法人，大学，さらには海外の政府関連組織までもが改ざん被害を受けていた．

4.2.3　1 次調査から得られた知見

調査結果を分析したところ，改ざんサイト及び不正スクリプトについて以下の 2 つの特徴が得られた．

(特徴 A) タイトルの改ざん

「タイトルの改ざん」とは，<title>タグ内に不正スクリプトが挿入されるという特徴である．この特徴は，調査した 5 割以上のサイトで確認できた．

(特徴 B) スクリプトの多重挿入

「スクリプトの多重挿入」とは，不正スクリプトが別の不正スクリプトによって上書きされている特徴のことである．多重挿入の例として"<<script~", "<scr<script~", "<scrip src=<script~"などが挙げられ，これ以外にも様々なパターンがある．

これら 2 つの特徴は，SQL インジェクション攻撃が無差別かつ自動的に行われることで必然的に生じる特徴であると考えられる．

4.3　2 次調査 (不正スクリプト収集)

4.3.1　調査目的と方法

2 次調査の目的は，4.2.3 で得られた特徴 A, B を用いて，3 章で述べたような「未知の不正スクリプト」が発見できるかを確かめることである．

特徴 A, B を用いる部分は，不正スクリプト収集時の検索キーワードである．2 次調査では，上記のキーワードによるサイト検索を行い，検索結果のサイトソースを分析することでスクリプトの抽出を行う(図 4).

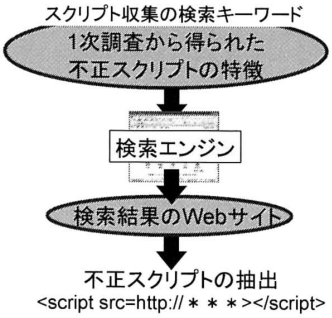

図 4　不正スクリプト収集方法

4.3.2　検索キーワード

「タイトルの改ざん」という特徴を有したサイトを検索するため，本調査では，ページタイトルのみを検索対象とする検索オプション " intitle: " を使用した(図 5,例 1)．また，「多重挿入」の特徴を有したサイトを検索するために，スクリプト重複の組み合わせ全てを検索キーワードとした(図 5,例 2).

例 1 : intitle:"<script src=http://*"
例 2 : "<s<script src=http://*"

図 5　検索キーワードの例

なお 2 次調査では，1 次調査のように特定の URL を指定した検索は行わず，総当り的な検索を行うため "http://～"以下を 'a'～'z', '0'～'9' とした．

4.3.3　調査結果

2 次調査では 243 件のスクリプトを収集した(表 2).

表 2　2 次調査の結果

既存ブラックリストとの比較	スクリプト(件)
掲載済	147
未掲載	96
合計	243

4.3.4　分析

2 次調査では，Shadowserver のブラックリスト[2]に掲載されていないスクリプトを，96 件発見した(表 2).

これらの未知スクリプトの信憑性を確かめるため，スクリプト中に記載された URL 先への接続実験を行っ

たところ，ボットやトロイの木馬などのマルウェアへの感染が確認された．この結果から，4.2.3 で示した特徴 A, B を用いることで，未知不正スクリプトは発見可能であるとわかった．また，検索エンジンを利用したスクリプトの収集法によって，効率的に未知の不正スクリプトを見つけることが期待できるとわかった．

また 1 次調査同様に，特徴についての分析を行ったところ，以下の 3 つの特徴を得ることができた．

(特徴 C) スクリプト名の偏り

スクリプト名とは，誘導先 URL "http:// ~ /?" の "?" の部分のことである．「スクリプト名の偏り」とは，このスクリプト名に図 6 に示すような偏りがあるという特徴である．なお，"単語 1 文字"とは"a.js", "1.js"などのことを指す．

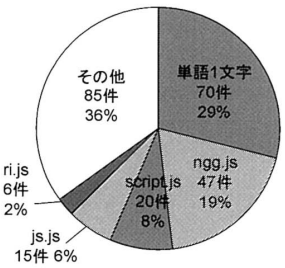

図 6　スクリプト名別の統計

(特徴 D) トップレベルドメインの偏り

「トップレベルドメインの偏り」とは，不正誘導先 URL のトップレベルドメインが".com", ".ru", ".cn"で 9 割以上を占めているという特徴である．

(特徴 E) スクリプト内の属性の設定・未設定

「スクリプト内の属性の設定・未設定」とは，特定の属性が設定もしくは未設定であるという特徴である．"type"や"language"といった属性は調査した全ての不正スクリプトで未設定だった．一方，"width", "height"などは，スクリプトを非表示にするため，いずれかが' 0 'に設定されていた．

5. 提案手法

提案手法は，4 章で得られた特徴を利用して，組織内ネットワークのプロキシにおいて改ざんサイトの判定・検知を行い，通信の遮断もしくは警告を行うシステムである．図 7 にそのシステム図を示す．

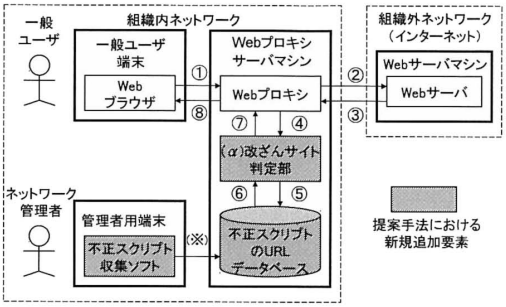

図 7　提案手法のシステム図

改ざんサイト判定部(図 7, α)では，4.2.3 の特徴 A,B 及び，4.3.4 の特徴 C,E を用いて考案した「特徴分析による判定」と，既存の「ブラックリストによる判定」によって改ざんサイトの判定を行うこととした．図 8 にその判定アルゴリズムを示す．

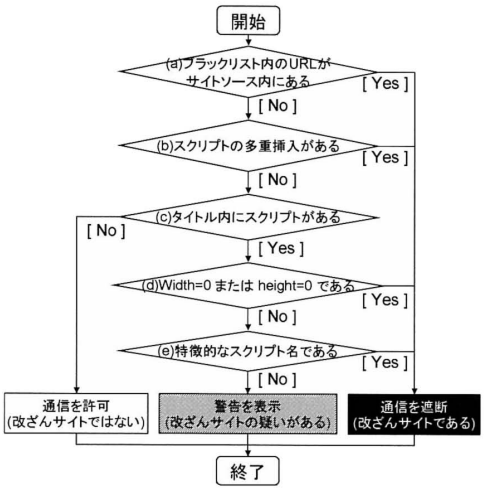

図 8　改ざんサイト判定アルゴリズム

ブラックリストによる判定は図 8,(a)で行われ，特徴分析による判定は図 8,(b)~(e)で行われる．アルゴリズムの作成にあたり，特徴 A,B(図 8,c,b)は，正常なサイトに存在する確率が極めて低く，なおかつ 4.2.3 で示したように，改ざん攻撃では必然的に起こりうる特徴であるという理由から上流に配置した．また特徴 C,E(図 8,e,d)は，正常なサイトにも存在する確率が高いと考えられることから，複数の特徴と組み合わせることが必

要であると考え，下流に配置した．

次に提案手法全体の処理の流れを以下に示す．

●データベースのメンテナンス時

1：定期的に管理者が不正スクリプト収集ソフトを動作させ，新しい不正スクリプトの URL を追加する(図 7,※)．ここで用いる収集ソフトのアルゴリズムは，4.3 の調査方法をベースとした検索エンジンによる収集アルゴリズムである．

●一般ユーザによる Web アクセス時

1：Web ブラウザからの Get 命令で，プロキシは対象URLのソースファイルを取得する(図7,①~③)．

2：取得したソースファイルについて，図 8 の判定アルゴリズムを用いて改ざんの有無をチェックする(図 7,④~⑦)．また，データベースに登録されていない不正スクリプトの URL を発見した場合は，データベースに追加する．

3：問題がなければブラウザで表示する(図 7,⑧)．改ざんが検知された場合は，通信を遮断，もしくは警告を表示する．

6. 考察

提案手法の「ブラックリスト収集方法」と「検知手法」について，既存手法との比較による考察を下記で述べる．

6.1 ブラックリスト収集方法の比較

提案手法と従来手法の比較を表 3 に示す．

表 3　ブラックリスト収集方法の比較

ブラックリストデータ収集方法	収集スピード	収集範囲
提案手法(能動的)	○	広い
従来手法(受動的)	×	狭い

提案手法の収集方法は，検索エンジンを用いて自動的かつ積極的に収集活動を行う方式である．そのため，報告をベースとした受身の従来手法に比べ，収集スピードが速く，収集可能範囲も広いといえる．実際，提案手法の収集方法によって，外部で公開された危険な誘導先ドメイン[3]を公開の一週間以上前に発見することができた．

6.2 検知手法の比較

ブラックリストによる判定の問題点であった未知の不正スクリプトに対しては，4.3.3 の表 2 で示したように検知・収集が可能であることが分かった．図 8 の判定アルゴリズムを用いて改ざんを検知した 20 件のサイトについて，セーフブラウジングによる警告の有無を確認したところ，5 件は警告が出ておらず，なおかつマルウェア感染の危険があることが確認された．また，この実験において判定の誤検知はなかった．

6.3 まとめ

6.1，6.2 より，既存の対策手法では見つけられなかった不正スクリプトを，提案手法では発見することができた．このことから，提案手法を用いることで，従来に比べ多くの改ざんサイトが検知可能になると期待できる．上記 6.2 の簡易実験においては，誤検知を確認することはなかったが，完全にその可能性がないとはいえず，今後，さらに多くのデータに基づいて評価を行っていく必要がある．

7. おわりに

本稿では，改ざんサイト及び不正スクリプトの調査・分析を通して，ユーザ標的型 Web 改ざん攻撃における特徴を発見し，そこから未知の不正スクリプトを自動的に検知可能な方法と，効果的な不正スクリプトの収集方法を考案した．そして，これらを用いて Web プロキシにおいて検知を行うシステムを対策手法として提案した．今後の課題は，iframe での改ざんに対する有効性の確認，誤検知率における実験評価，および提案手法の実装・評価を行うことである．

参考文献

[1] LAC，侵入傾向分析レポート Vol.11, 2008.09.17

[2] Shadowserver Foundation , http://www.shadowserver.org/wiki/pmwiki.php?n=Calendar.20080514 (2009.01.22)

[3] LAC，【緊急注意喚起】改ざんされた Web サイト閲覧による組織内へのボット潜入被害について，http://www.lac.co.jp/news/press20081222.html (2009.01.22)