社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

不正サイト検知のための自律分散型 Web クローラと仮想化基盤の提案

河野 義広† 三須 剛史‡ 花田 真樹‡ 布広 永示‡

†東京情報大学 〒265-8501 千葉県千葉市若葉区御成台 4-1

E-mail: †ykawano@rsch.tuis.ac.jp, s11118tm@gmail.com, mhanada@rsch.tuis.ac.jp, nunohiro@rsch.tuis.ac.jp

あらまし 本研究では，Web 上の不正サイトの効率的な検知を目的とした自律分散型 Web クローラの開発を行う．本システムは，自律分散型 Web クローラとその実行基盤により実現される．具体的には，Web サイトの短縮 URL のハッシュ値を起点として，複数の Web クローラが並列に探索を行う．加えて，各クローラを仮想化基盤上で実行することで，不正サイト検知時の影響範囲を限定し，安全かつ効率的な探索を実現する．

キーワード サイバーセキュリティ，Web クローラ，自律分散協調システム，仮想化

A Proposal of Distributed Autonomous Web Crawler and Virtualization Infrastructure for Detection of Malicious Web sites

Yoshihiro KAWANO† Takeshi MISU, Masaki HANADA, and Eiji NUNOHIRO‡

†Tokyo University of Information Sciences, 4-1 Onaridai, Wakaba-ku, Chiba, 265-8501 Japan

‡Freelance

E-mail: †ykawano@rsch.tuis.ac.jp, ‡yuka.obu@gmail.com

Abstract Recently, cyber security to protect information systems or personal information against threats of the internet is big issue. We are studying a distributed autonomous cooperative system about exclusive Web crawling for detection of malicious Web sites. The system is divided into three subsystems, that is, Web space projection system, autonomous cooperative Web crawler, and the crawler runtime environment. The Web crawlers to collect Web sites based on URL shortening effectively and exclusively. In addition, by introduction of virtualization technology to crawler runtime environment, dynamic reconstruction for effective detection of malicious Web sites is realized. Future works are development of the crawler and the runtime environment, and evaluation.

Keyword Cyber security, Web crawler, Distributed Autonomous Cooperative System, Virtualization

1. はじめに

近年，マルウェア感染やフィッシングサイト，なりすましなど，Web を介したサイバーセキュリティの脅威が社会問題となっている．例えば，マルウェアの感染経路として，企業や個人，自治体などの Web サイトを改ざんし，マルウェア配布先の Web サイトに誘導するドライブ・バイ・ダウンロード攻撃による被害の報告されている[1]．加えて，短縮 URL サービスを用いてマルウェアやワンクリック詐欺などの不正サイトを隠蔽し，Twitter や Facebook をはじめソーシャルメディア上から拡散する手法も増加している[2]．このような不正サイトは，手法を変えながら日々増加しており，セキュリティに対する意識や知識を十分に持たない一般的な生活者が Web の危険性を認識しながら利用することは難しい．

そこで本研究では，不正サイトを効率的に検知することを目的とし，排他的な Web クローリングのための自律分散協調システムを開発する．具体例には，広大な Web 空間に点在する不正サイトを効率的に収集するため，Web 空間をセキュリティリスクの特徴量に基づく N 次元空間に射影し，その空間を排他的に探索する自律型 Web クローラの基盤技術を研究する．各 Web クローラは，射影空間内での座標や探索方向などを共有し，並列分散処理技術を応用することで探索空間を独自に判断する．加えて，仮想化技術を導入することで，不正サイトの特性や検知目的に対応した自律分散協調システムの実行環境を動的に再構築し，サイバーセキュリティ対策に係わる解析精度を向上する．

2. 提案システム

2.1. システム概要

提案システムは，Web 空間射影システム，自律分散協調 Web クローラ，クローラ仮想化マネージャの 3 システムに分類できる．各システムの特徴を以下に示す．

1. Web 空間射影システム
   - 目的：各クローラが Web サイトを効率的に探索するための分類方法を見つけること

※ 1 クローラに負荷が集中したり，一斉に同じ箇所を探索したりしないようにする

- 要素技術：セキュリティ，多変量解析

2. 自律分散協調 Web クローラ
   - 目的：複数のクローラが協力して効率的・排他的に Web 空間を探索できること
   - 要素技術：並列分散処理，協調アルゴリズム
3. 仮想化基盤
   - 目的：クローラが不正サイトを発見した場合，迅速に実行環境を再構築できること
   - 要素技術：仮想化技術

上記テーマはそれぞれ独立に研究可能であり，それらの成果物が統合されることで自律分散協調システムが完成する．

2.2. Web 空間射影システム

Web 空間を探索可能な空間にマッピングする際の必要条件は以下のとおりである．下記の条件を満たすことができれば，各 Web サイトは空間上の点で表現されるため，各 2 点間の距離が算出できる．

(1) Web サイトの所在を座標で表現できること
※幾何学性質が利用できる 2，3 次元が適切

(2) 座標の重複がないこと
※同一座標に複数の Web サイトがマッピングされないこと

(3) 空間的に均等に配分されること
※特定の領域に点が集中しないこと

(4) 上記が一意に定まること
※同一のアルゴリズムを使用し，各クローラが独自に判断できること

上記手順において，短縮 URL とそのドメインの whois 登録日時を 2 軸として，射影空間を算出する（図 1）．短縮 URL とは，提供サービスのドメインと 6～10 文字程度のランダムな英数字の組み合わせで構成される URL であり，ビジネスや教育現場での Web 資料や Twitter 内でのリンク共有に利用されることが多い．近年，短縮 URL サービスは不正サイト隠蔽の温床として利用される事例が報告されていることから，短縮 URL サービス毎にハッシュ値の文字列組み合わせを順次走査しながらクローリングを行う．収集した Web サイト情報の蓄積に関して，扱うデータが大規模になること，不正サイト検出時に実行環境の再構築が必要となることから，収集・解析の処理をリアルタイムかつ高速に実行しなければならない．Web 空間を 2 次元の軸に射影する手法を以下に示す．

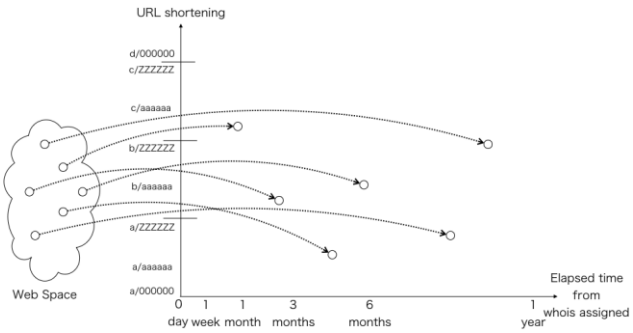

図 1. Web 空間の射影

＜x 軸：whois 登録期間＞

(i) ドメインの whois 登録日からの経過日数を計算
(ii) 経過日数が 1 年未満なら手順(iii)，それ以外なら空間上に射影せずに終了
(iii) 経過日数の二進対数を x 軸の値とする

＜y 軸：短縮 URL＞

(i) 短縮 URL サービスの識別子とハッシュ値の組み合わせで 11 桁の英数字を算出する
(ii) (i)の値を昇順にして y 軸の値とする

上記の x 軸において，ドメインの登録期間が短いサイトは，不正サイトの危険性が高いものと想定する．x 軸を二進対数としたのは，登録期間が短いほうがより危険性が高いと考えたためである．次に y 軸は，短縮 URL サービス毎にハッシュ値を昇順で並べて配置する．採用する短縮 URL サービスは，不正サイトの温床となっているものを中心に選別する．

2.3. 自律分散協調 Web クローラ

前節にて空間内に射影した Web サイトを自律的に収集する Web クローラを開発する．各クローラの実行手順は以下のとおりである．

(1) DB の Web サイト情報をもとに，探索空間を独自に決定する
(2) 探索空間内の Web サイトを収集し，不正サイトか解析する
(3) 解析結果を DB に登録する
(4) (1)に戻り探索空間を再計算する
※他クローラの解析状況に基づき計算する

(1)にて，空間内における各クローラの探索位置や探索方向をもとに，ボロノイ図，もしくはマハラノビス距離を用いて探索空間を各クローラが独自に決定する．提案手法は，文献[3]を参考に研究を進める．(2)にて，

取得した Web サイトの HTML や JavaScript などを解析し，不正サイト（例えば，フィッシングやなりすましサイト）の判断を行う．(3)にて，Web サイトの解析結果を DB に登録することで，他クローラと解析状況を共有する．(4)では，探索空間を再計算し，(1)～(3)の処理を繰り返す．各クローラの起動・停止の制御は，以降で説明する仮想化マネージャにより実現する．

＜探索領域決定方法＞
前提条件：
- クローラの単位時間の移動距離を制限
  ※解析学の性質（連続性）を利用するため
- クローラ同士は直接通信しない
  DB を介して、他のクローラの状況把握

手順：
(i) DB より隣接クローラの探索位置と移動履歴を取得
(ii) (i)に基づきマハラノビス距離を計算し自クローラの探索領域を決定
(iii) 移動可能な点から、探索時期が 1 週間以上前かつ自クローラの探索領域を最大化する点を選択

2.4. クローラ仮想化マネージャ

2.3 節にて開発した Web クローラの実行環境を開発する．仮想化技術を用いて，各クローラの起動・停止を制御するクローラ仮想化マネージャ（Crawler Virtualization Manger；以下 CVM）を開発する．システムアーキテクチャを図 2 に示す．図 2 では，仮想化基盤上のシステム構成であり，CVM はハイパーバイザ上に構築されたゲスト OS の 1 つとして実装する．CVM の必要条件は以下のとおりである．

(1) 解析状況に応じて、クローラの追加・削除ができること ※空間の探索効率化
(2) 不正サイト検出時に、当該クローラを初期化できること ※クローラの安全性の確保
(3) クローラが他のプロセスに干渉できないこと ※実行環境の安全性の確保

(1)では，DB の解析状況をもとに，クローラの追加・削除の判断を行い，それを実行できる機能が必要となる．(2)では，DB の解析をもとに不正サイト検出の有無を判断し，当該クローラの停止・再起動の処理を行う．(3)では，クローラがシステム内の他のプロセスやネットワークに影響を与えないよう，実行環境を隔離する必要がある．

そこで本研究では，CVM の要素技術としてコンテナ型仮想化技術「Docker」を採用する[4]．Docker は，Xen や KVM などの従来のハイパーバイザ型仮想化技術と比較して，特定のプロセスを軽量かつ高速に実行できる仮想化技術である．Docker コンテナに格納されたプロセスは，コンテナ外のプロセス ID の参照や通信ができない仕組みとなっているため，各クローラの安全かつ高速な実行環境を実現できる．一方，ホスト OS 側からは Docker コンテナ内のプロセス ID を参照できるため，DB の解析状況をもとに，各クローラの動的再構築が可能となる．CVM の手順を以下に示す．

＜CVM の手順＞
(i) DB を参照し、クローラの起動・停止を判断
(ii) (i)に基づき Docker 構成ファイルの更新
(iii) Docker コマンドによりクローラの動作を制御

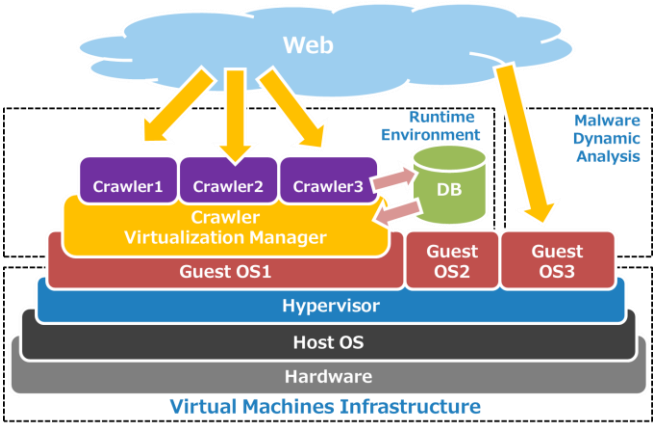

図 2. 提案システムの全体構成

3. システム設計

3.1. 概要

システムデータフローを図 3 に示す．図 3 において，本システムは Web クローラと CVM に分割できる．Web クローラは DB の状態をもとに探索領域を独自に決定する．各クローラから DB へのアクセスは，DB 用の API を介して実行する．

Web クローラは，探索のエントリポイントとなるアドレスを短縮 URL の規則をもとに自動生成し，展開後のドメインの whois を確認する．その後，URL 展開後の Web サイトの HTML や JavaScript を取得し，不正サイトかどうかの判定を行う．判定処理については，筆者らが研究した手法を用いる[5]．この手法では，攻撃者によって改ざんされた Web サイト内のコードを比較し，オンライン URL スキャンサイト「VirulTotal[6]」を用いて不正サイトの判別を行う．

一方，CVM は DB に記録された解析結果を定期的に確認し，当該クローラが不正サイトにアクセスした場合に即座にそのクローラの再起動を実行する．

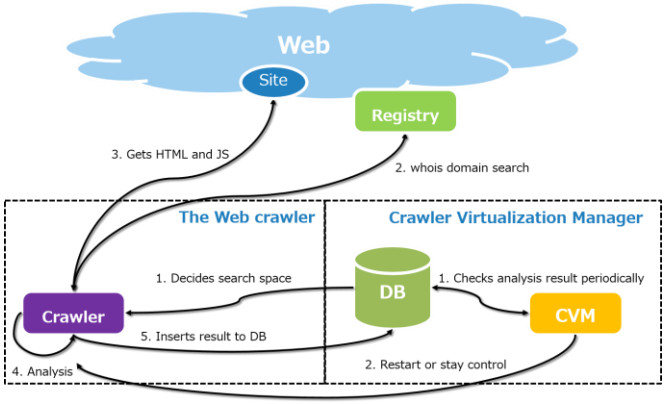

図 3. システムデータフロー

3.2. DB 設計

本システムの ER 図とテーブル定義表を図 4，表 1-3 に示す．本システムでは，Web サイト（web_sites），クローラ（crawlers），解析結果（analysis_logs）の 3 つのテーブルを利用する．Web サイトテーブルでは主に短縮 URL と whois 登録日時，クローラテーブルでは Docker コンテナ ID と探索回数や探索位置の座標，解析結果テーブルでは探索したサイトのフィッシングやなりすましなどの評価値を，それぞれ記録する．

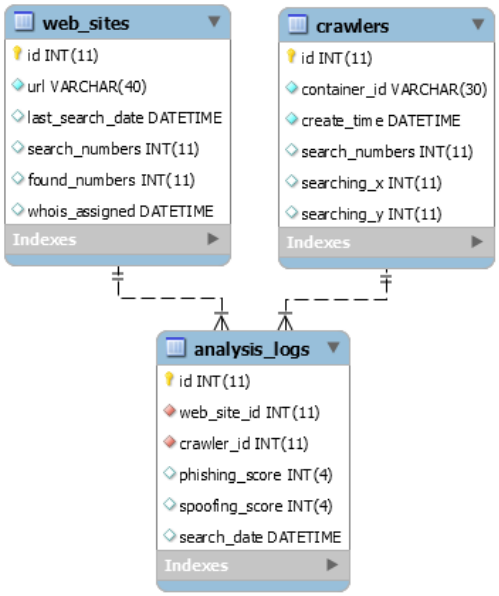

図 4. DB の ER 図

表 1. Web サイトテーブルの定義表

フィールド名	型	Key	意味
id	INT(11)	PRI	ID
url	VARCHAR (40)		短縮 URL
last_search_date	DATETIME		最終探索日時
search_numbers	INT(11)		探索回数
found_numbers	INT(11)		不正サイトの発見数
whois_assigned	DATETIME		Whois 登録日

表 2. クローラテーブルの定義表

フィールド名	型	Key	意味
id	INT(11)	PRI	ID
container_id	VARCHAR (30)		Dokcer コンテナ ID
create_time	DATETIME		Docker インスタンスの作成日時
search_numbers	INT(11)		探索回数
searching_x	INT(11)		x 座標の位置
searching_y	INT(11)		y 座標の位置

表 3. 解析結果テーブルの定義表

フィールド名	型	Key	意味
id	INT(11)	PRI	ID
web_site_id	INT(11)	MUL	Web サイトテーブルの ID
crawler_id	INT(11)	MUL	クローラテーブルの ID
phishing_score	INT(11)		フィッシングの評価値
spoofing_score	INT(11)		なりすましの評価値
search_date	DATETIME		探索日時

4. まとめ

本稿では，不正サイト検知のための自律分散型 Web クローラと仮想化基盤の提案を行った．具体的には，Web サイトの短縮 URL のハッシュ値を起点として，複数の Web クローラが並列に探索を行う手法を提案した．加えて，各クローラを仮想化基盤上で実行することで，不正サイト検知時の安全性と効率性を実現する．

今後の課題は，本システムのプロトタイプ開発，有効性の検証方法の検討などが挙げられる．

文　献

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

[2] Intel Securit，“Short-URL Services May Hide Threats”，McAfee Securing Tomorrow Today，https://securingtomorrow.mcafee.com/mcafee-labs/short-url-services-may-hide-threats/，2013.7.

[3] T. Yonekura, Y. Kawano, "A Protocol for Peer-to-peer Multi-Player Networked Virtual Ball Game", IEICE Trans. INF. &SYST., Vol.E88-D, No.5, pp.926-937, 2005.5.

[4] Docker, https://www.docker.com/.

[5] 三須剛史，佐藤順子，花田真樹，山口崇志，布広永示，”セキュリティインシデント解析支援を目的とした悪性 Web サイト発見システムの提案”，コンピュータセキュリティシンポジウム 2016 (CSS2016)，2016.10.

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