Vol.2017-CSEC-78 No.5
Vol.2017-SPT-24 No.5
2017/7/14

サイト関連情報に基づいた Web サイト脅威度推定機能の提案

藤井翔太[†1]　鬼頭哲郎[†1]　重本倫宏[†1]　藤井康広[†1]

概要：サイバー攻撃の激化に伴い，悪性 Web サイトへのアクセスを防止する技術が求められている．このような背景から，我々は，不審な Web サイトへのアクセスに対して，マルウェアでは突破困難な追加認証を課す技術を開発しているが，無害な Web サイトに対しても追加認証を課してしまう場合があり，業務への悪影響となりうる．そこで，本稿では，サイト関連情報に基づいた Web サイト脅威度推定機能を提案する．本機能は，DNS レコード等のサイト関連情報を既知のブラック/ホワイトリストから学習し，Web サイトの脅威度を推定する．推定結果を用いて，脅威度の低い Web サイトへの追加認証を抑制することにより，業務への悪影響を低減する．

Proposal of Website Risk Estimation Function based on Website Related Information

SHOTA FUJII[†1]　TETSURO KITO[†1]
TOMOHIRO SHIGEMOTO[†1]　YASUHIRO FUJII[†1]

Abstract: Along with the increasing of cyber attacks, preventing the spread of its damage is required. In this situation, we have proposed AED, which prevents the access to malicious websites using additional authentication. However, there is a possible that AED imposes additional authentication on harmless websites and it may cause the reducing work efficiency. To resolve this problem, we propose website risk estimation function. This function learns features of malicious/benign websites beforehand and estimates the risk of websites based on beforehand learning. This function reduces the negative impact of AED on work efficiency by suppressing additional authentication for low risk websites using estimation result.

1. はじめに

近年，標的型攻撃に見られるように，攻撃が高度化しており，企業や国家にとって重大な脅威となっている．また，攻撃の高度化に伴い，外部からの侵入を完全に遮断することは困難になってきている．侵入された後の被害を抑制するためには，機密情報の窃取や感染拡大を目的とした悪意のある Web サイト（以降，悪性サイト）等，外部への通信を遮断することが重要である．ここで，悪性サイトへの通信を抑制する方法のひとつに，インテリジェンス（ブラックリスト等）を用いるものがある．しかし，インテリジェンスには，潜在的に偽陽性が含まれており，仮に業務遂行に必要な非悪性サイトがブラックリストに誤って含まれていた場合，当該サイトにアクセスできず，業務阻害の要因となってしまう．業務阻害を抑制する方法として，インテリジェンスを事前に精査し，非悪性サイトを人手で除外する方法も考えられるが，Web サイトの精査という別のコストが生じてしまう．

このような状況を受け，我々は，「自律進化型防御技術 (Autonomous Evolution of Defense，以降 AED)」の研究を進めている[1][2]．AED は，インテリジェンスに含まれる不審 URL をグレーリストという形で保持し，一旦グレーなものとして保留する．その後，グレーなサイトへのアクセスを検知した際，追加認証を要求する．同アクセスが人間によるものだった場合は，文字を読み取り，追加認証を突破することでアクセスを続行できるが，マルウェアは文字を認識できず，追加認証を突破できないため，アクセスを防止できる．このように，不確実なインテリジェンスであっても，業務への悪影響を最小限に抑制しつつ，対策への活用が可能となる．さらに，複数の利用者が追加認証を突破した場合，当該サイトは安全なものとしてホワイトリストへ振り分けるといったように，追加認証結果に応じて，リストを自動生成する機能も有する．これにより，従来のブラック/ホワイトリスト方式の課題であったリスト管理のコストを抑制可能である．

一方で，AED には，膨大なインテリジェンスを利用した場合，追加認証が頻発し得る問題，および，追加認証の結果が望ましくないものであった場合，リストが汚染される可能性がある問題が残されている．そこで，本稿では，両問題を解消する Web サイト脅威度推定機能の設計を述べる．また，設計に基づいてプロトタイプの実装を行い，評価した結果を述べる．

2. 自律進化型防御技術（AED）

2.1 AED の全体像

AED とは，不審 URL へのアクセス制御を，業務への悪影響を抑えつつ迅速に行うための技術である．AED は，いきなり通信を止めるのではなく，一旦グレーなものとして保留する．そして，グレーなサイトにアクセスしたとき，追加認証を行う．人間はアクセス先が悪性でないと判断した場合，文字を読み取り，追加認証を突破して外部にアクセスできるが，マルウェアは文字を認識できず追加認証を

†1 株式会社日立製作所
Hitachi Ltd.

Vol.2017-CSEC-78 No.5
Vol.2017-SPT-24 No.5
2017/7/14

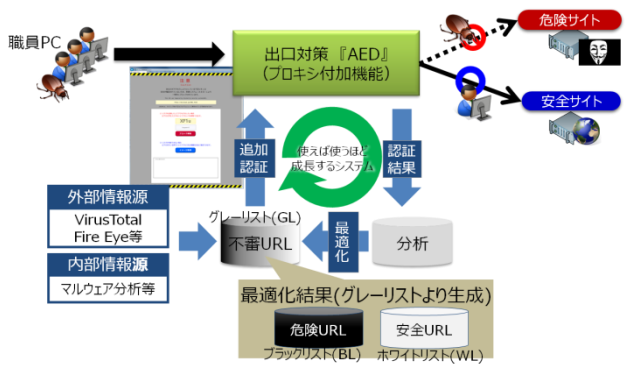

図 1　AED の全体像

突破できないため，アクセスを引き続き防止可能である．なお，アクセス先が悪性か否かの判断を補助するために，ユーザの要求に応じてアクセス先のスクリーンショット等を取得する機能も有する．この追加認証が AED の第一の特徴である．そして，複数の従業員が追加認証を突破した場合，安全なサイトとみなしてグレーリストを自動でホワイトリストに振り分ける．誰も追加認証を突破できなかった場合，悪性なサイトだとみなしてブラックリストに振り分ける．このように，従業員の追加認証結果を統計的に分析して，アクセスポリシが自律進化していくことが AED の第二の特徴である．

以上のように，AED を用いれば，ブラックリストを事前に精査せずとも自動で更新されるため，業務への悪影響がなく，かつ迅速に対処でき，運用を効率化することが可能となる（図 1）．

2.2 追加認証に係る残された課題

2.1 節で述べたように，AED は，追加認証を用いることによって，業務への悪影響を抑制しつつ，悪性サイトへのアクセスを防止可能である．一方で，この追加認証に関して，以下の残された課題がある．

(課題1) 追加認証が頻発し得る

AED は，不審なサイトを一旦グレーリストに振り分けるため，膨大な外部インテリジェンスを利用した場合，グレーリストが膨大になり得る．また，リストにない悪性サイトへのアクセスを抑制したい場合は，ブラックリストとホワイトリストのいずれにも含まれないサイトは全てグレーとして扱い，ユーザに追加認証を要求する必要がある．この際は，上記の場合よりもさらにグレーリストが肥大してしまう．これらの場合，肥大したグレーリストへのアクセスに対して追加認証を要求するため，ユーザの追加認証数が増加し，ユーザビリティの低下に繋がる恐れがある．

(課題2) 認証結果が望ましいものではない可能性がある

既存技術では，認証結果の精度については考慮されていない．このため，ユーザの知識不足や操作ミスにより，追加認証の結果が本来あるべきものと異なってしまった場合，グレーリストからの各リストへの振り分け精度が低下し，脅威度の高いサイトへのアクセスが発生してしまう可能性がある．

また，追加認証結果には，利用者の行動結果が反映されている．このようなデータは，機械的な生成が難しく，比較的貴重なものであることから，他の用途への転用が期待できる．ただし，追加認証結果の中に望ましくないものが混入していた場合，転用先でも悪影響が生じる可能性がある．

以降では，上記の課題に対処する方式を述べる．

3. Web サイト脅威度推定機能

3.1 要件と実現のための機能

2.2 節の各課題を解決するための要件は，以下の 2 つである．

(要件1) 高精度なアクセス先 Web サイトの精査

（課題 1）への対処として，高精度にアクセス先 Web サイトを精査する機能の実現がある．本機能を用いて事前にアクセス先 Web サイトの脅威度を推定し，脅威度の低いものに対しては追加認証無しでアクセスを許可することにより，不要な追加認証を抑制できる．

(要件2) 望ましくない認証結果の識別

（課題 2）への対処として，望ましくない認証結果を識別することがある．この際，望ましくない認証結果は排除することにより，精度低下の抑制が可能になる．

各要件を満足する機能を検討し，以下の 3 つを導出した．

(機能1) 複数推定器を備えた脅威度推定機能

以降では，機械学習を用いて対象の性質を推定する機構を推定器と呼ぶ．ここで，一般的に複数の推定器を組み合わせることにより，その推定精度が向上することが知られている．そこで，（要件 1）を満足するため，Web サイト脅威度推定機能は，複数の推定器を備えた脅威度推定機能を実現する．

また，全体のうち，精度がより高い推定器の推定結果を重用することで，全体としての推定精度も向上できると考えられる．ここで，AED は，脅威度推定器によって脅威度が高いと判定されたサイトに対しては，利用者に追加認証を要求し，その正否によってアクセス可否を判断する．この利用者の追加認証結果と脅威度推定機能を構成する個々の推定器の推定結果を突合し，その両者が一致したものは精度が高いものとして重用することで，前述の推定精度向上が期待できる．そこで，利用者の追加認証結果を基に，各推定器に対して適切に重みを与える機能も併せて実現する．

(機能2) 外部情報取得機能

Web サイト脅威度推定機能は，後述のように，アクセス先サイトの脅威度を推定するために，URL 文字列以外

の情報も利用する．この情報として，例えば，WHOIS 情報や DNS 情報のように，当該サイト以外の外部サイトから取得する必要のあるものがある．そこで，URL を基に，外部サイトから取得する必要のある情報（以降，外部情報）を取得する．

(機能3) 認証結果検証機能

（要件 2）を満足するために，認証結果検証機能を実現する．本機能は，認証結果が望ましくないものか否かを識別する機能である．この機能を実現することにより，前述したリスト振り分け機能の精度低下を抑制するとともに，（機能 1）の項で述べた利用者の追加認証結果に基づく推定器への重み割り振りを適切に実施する．

以降の節では，各機能の詳細とその実現方法を述べる．

3.2 各機能の設計

3.2.1 複数推定器を備えた脅威度推定機能

Web サイト脅威度推定機能は，3.1 節で述べたように，複数の脅威度推定器を利用し，アクセス先の最終的な脅威度を予測する．今回は，3 つの推定器を用いることを前提とする．このとき，推定器 1 が URL 文字列，推定器 2 が WHOIS 情報，および推定器 3 がその他の情報（DNS 情報，地理情報，および Alexa ランク）を利用してアクセス先サイトの脅威度を判定する．

また，上記の情報は，良性サイトと悪性サイトの間に違いとして現れやすい点に着目して選出している．選定した特徴量を表 1 に示す．なお、各特長量の値は，0〜1 の間に正規化して利用する。

表 1　各推定器における特徴量と値域

推測器	カテゴリ	通番	特徴量	値
1	URL 文字列	1	URL文字列長	0-
		2	ドメイン文字列長	0-
		3	パス文字列長	0-
		4	URL文字列に含まれる数字の数	0-
		5	ドメイン文字列に含まれる数字の数	0-
		6	パス文字列に含まれる数字の数	0-
		7	パス文字列に含まれるトークン数	0-
		8	パス文字列に含まれる平均トークン数	0-
		9	パス文字列に含まれる最長トークン数	0-
		10	ドメイン文字列全体に対する最長英単語の長さが占める割合	0-1
		11	FQDNのジニ係数	0-1
		12	FQDNが“.”を含むか	0/1
		13	FQDNが“-”を含むか	0/1
		14	URLが拡張子で終わるか否か	0/1
		15	URL文字列が'exe'を含むか否か	0/1
		16	URL文字列が'php'を含むか否か	0/1
		17	ドメインがIPアドレスか否か	0/1
2	WHOIS 情報	18	ドメイン登録期間（年数）	0-10
		19	ドメイン性質（初登録or更新有）	0/1
		20	ドメイン初登録からの年数	0-
		21	ドメイン登録時間（0-23）	0/1
		22	ドメイン登録曜日（月火水木金土日）	0/1
		23	レジストラ（当該レジストラのうち，悪性データ分が占める割合）	0-1
3	DNS 情報	24	Aレコード数	0-
		25	AAAAレコード数	0-
		26	CNAMEレコード数	0-
		27	MXレコード数	0-
		28	NSレコード数	0-
		29	PTRレコード数	0-
		30	TXTレコード数	0-
		31	逆引きが設定されているか否か	0/1
		32	ネガティブTTL（SOAレコードのminimum値）	0-
	Alexa ランク	33	Alexaランク（訪問者数ランク）	0-
		34	Alexaランクの差分（アクセス数ランク−訪問者数ランク）	0-
	地理情報	35	国（IPアドレスに対応する国のうち，悪性データ分が占める割合）	0-1

URL 文字列では，悪性 URL は，良性 URL よりも，複雑である[3][4][5]（通番 1〜11），特定の文字・拡張子の出現頻度が高い[3][5][6]（通番 12〜16），およびドメインと紐付けられていない[4][5]（通番 17）という傾向を掴むために特徴量を選定した．なお，通番 11 のジニ係数とは，集合の複雑性を測る指標の 1 つであり，集合の複雑性が低いほど 0 に，高いほど 1 に近づく．このため，FQDN をアルファベットや記号 1 文字ごとの集合とみなし，ジニ係数を算出することにより，その複雑性を測ることができる．

WHOIS 情報では，攻撃に利用されるドメインは使い捨てであり，生存期間が短い場合がある[7][8][9][10]（通番 18〜20），攻撃者の登録作業コストや登録の金額面でのコストを抑制するために，まとめて登録される場合がある[3][8]（通番 21，22），および，攻撃者がドメインを取得する際のレジストラには偏りがある[7][8][11][12]という傾向を掴むために，通番 23 を選定した．

DNS 情報では，各種レコード（通番 24〜30）に加え，悪性サイトのネガティブ TTL は短い傾向にある[13][14]という特徴を掴むために，特徴量（通番 31）を選定した．

また，悪性サイトは，アクセス数が正規サイトより少ないという仮定の下，アクセス数ランキングである Alexa ランク（通番 32，33）を利用した．さらに．攻撃に利用されるドメインは特定の国に偏っているとの報告[6]から，通番 34 を選出した．

また，各推定器に与える重みには，学習時に教師データの分類精度を記録しておき，その分類精度の推定器間での比率を正規化した値を利用する．例えば，同様の教師データを推定器 1 が 90%，推定器 2 が 85%，推定器 3 が 75%の精度で分類できた場合，その比率は，90:85:75 であり，この比率の合計が 1 になるよう正規化する．つまり，それぞれの推定器に対して，0.36，0.34，0.30 の重みを与える．さらに，新たにアクセス先サイトの脅威度を推定した際は，各推定器の推定結果と利用者の追加認証結果を突合し，両者が一致したものの重みを大きくし，一致しなかったものの重みを小さくする．これにより，各推定器に与える重みを最適化する．

3.2.2 外部情報取得機能

本機能は，脅威度推定器を構成する各学習器が利用する情報に応じて，適宜実現する．例えば，3.2.1 項で挙げた推定器群を用いる場合，推定対象 Web サイトの URL をベースとして，外部のサーバに問い合わせを行い，WHOIS 情報，DNS 情報，Alexa ランク，および地理情報を取得する．

また，一度取得した情報は，DB 等に保管することにより，以降外部アクセスなしに参照できる．このため，同じドメインに対する 2 回目以降のアクセスでは，DB を参照して外部情報を取得することにより，外部アクセスを伴うことによる処理時間の長大化や同一情報提供元に対するアクセス過多を抑制できる．

Vol.2017-CSEC-78 No.5
Vol.2017-SPT-24 No.5
2017/7/14

3.2.3 認証結果検証機能

既存の AED は，2.2 節で述べたように，利用者の認証結果を検証しておらず，ユーザの知識不足や操作ミスにより，追加認証の結果が本来あるべきものと異なってしまった場合，グレーリストからの各リストへの振り分け精度が低下し，脅威度が高いサイトへのアクセスが発生してしまう可能性がある．また，本機能でも，（機能 1）において，利用者の認証結果を各推定器の重み割り振りに利用しており，望ましくない追加認証結果による悪影響を受ける可能性がある．そこで，本機能によって，認証結果を精査し，その確からしさが高いもののみを採用することで，上記問題を抑制する．このために，以下に示す 3 つの情報を利用する．

(1) 認証に要した時間

認証に要した時間は，主に人間によるアクセスとマルウェアによるアクセスを分別するために用いる．人間は追加認証が出た際に人間は柔軟に対応できるのに対し，追加認証を想定していないマルウェアは，これを突破できずタイムアウトになる．あるいは，近年の攻撃の高度化に伴い，追加認証を突破する機能を持つマルウェアも報告されているが，この場合は，プログラムによって人間では達成困難な入力速度で認証突破を試みると考えられる．以上のように，マルウェアが認証に要する時間には，タイムアウト，あるいはきわめて短いといった特徴があり，人間による認証と分類に有用であることから，本情報を利用する．

(2) 正解数

正解数とは，CAPTCHA 追加認証突破に必要な文字列のうち，どれだけ正解したかの数である．この正解数は，主に認証失敗時，該失敗がヒューマンエラーによるものか否かを分別するために用いる．例えば，認証に失敗した際，意図したものであれば，何も入力しない，あるいは適当な文字を入力することにより，認証突破のために要求された文字列から大きく外れることが考えられる．一方で，意図しないもの（ヒューマンエラー）は，突破しようとしたものの外れてしまった入力であることから，認証突破のために要求された文字列から大きくは外れないことが考えられる．以上のように，認証の失敗がヒューマンエラーによるか否かは，認証突破のために要求された文字列に対する入力の正解数に現れることから，本情報を利用する．

(3) サイト情報表示有無

前述の通り，AED は，不審なサイトへのアクセスを検出した際，一旦アクセスを保留し，追加認証を提示してアクセスを続行しても良いかの判断を利用者に委ねる．この際，利用者にとっては，URL からだけではアクセスを続行しても良いか否かを判断するのが難しい場合がある．この様な場合に対応するため，既存の

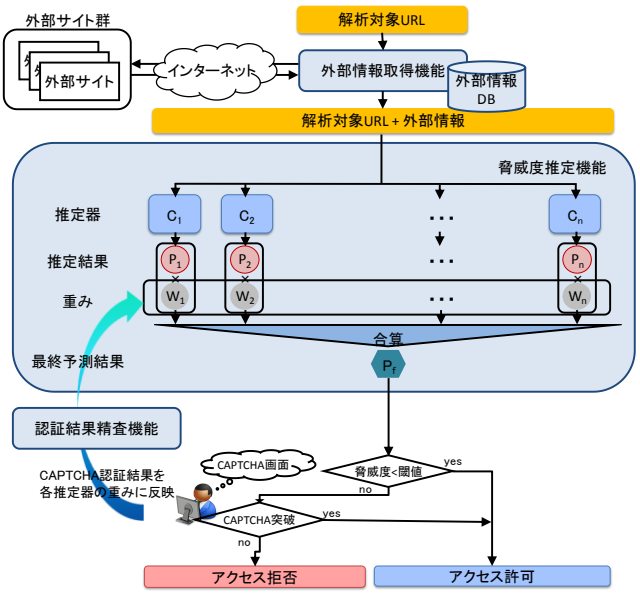

図 2　Web サイト脅威度推定機能の全体像

AED は，ユーザの要求に応じて，実際にアクセスする前にアクセス先 Web サイトの情報を確認するための機能を有する．具体的には，Web サイトのスクリーンショット等を追加で取得・表示する機能がある．これにより，ユーザのアクセス続行判断を補助する．ここで，当該情報を要求したユーザは，アクセス先 Web サイトの性質を多角的に判断しようとするセキュリティ意識の高いユーザであることが考えられることに加え，その際の判断は追加情報を見た上のものであることから，追加情報を見ていない判断よりも信頼性が高いと考えられる．以上のように，サイト情報表示有無は，ユーザの認証結果の信頼性を検証するのに有用であることから，本情報を利用する．

本機能により，認証結果の確からしさを数値化し，その確からしさに応じて Web サイト脅威度推定機能を構成する各推定器へフィードバックを行うことが可能となる．

3.3 Web サイト脅威度推定機能の全体像

Web サイト脅威度推定機能は，3.2 節で述べた機能と既存の AED を組み合わせて実現されるものであり，3.1 節で述べた 2 つの要件を満足する．図 2 に本機構の全体像を示し，以下でアクセス先をリアルタイムで解析，アクセス制御を行う場合の処理を説明する．

まず，外部情報取得機能を用いて，外部サイトから解析対象 URL に関する情報を取得する．その後，解析対象 URL と取得した外部情報を脅威度推定機能に投入し，アクセス先の脅威度を算出する．当該脅威度が事前に設定した閾値よりも高い場合，利用者に追加認証を提示し，その突破可否でアクセスを許可/拒否する．また，このときの認証情報を認証結果精査機能にかけ，確からしい認証であった場合，認証結果と各推定器の結果を突合した後，両者が一致しているか否かで，各推測器に係る重みを増減させる．

表 2　各外部情報の取得方法

カテゴリ	取得方法
WHOIS Information	Python-whois [16]
DNS Information	dnspython [17]
Alexa Rank	Alexa API [18]
Geological Information	GeoIP [19]

4. 評価

今回は，3 章で述べた 3 機能のうち，（機能 1）複数学習器を備えた脅威度推定機能と（機能 2）外部情報取得機能の 2 つを実装し，評価を行った．

4.1 実装

両機能を実現するにあたって，プログラミング言語 python を用いた．また，（機能 1）は，機械学習によってサイトの脅威度を推定するが，この部分の実装には機械学習ライブラリである scikit-learn[15]を利用した．（機能 2）での情報取得には，表 2 に示すライブラリ・情報元を利用した．なお，URL 文字列に関しては，外部情報を参照する必要が無いため，それ以外の情報について記載している．

なお，（機能 3）に関しては，後述する理由により，今回の評価実施が困難なことから実装を見送っている．

4.2 評価項目

本節では，評価項目を述べる．今回は，（要件 1），および当該の要件に対応する（機能 1)と（機能 2）を評価対象としている．評価項目は，以下の通りであり，(1)が（機能 2），(2)と(3)が（機能 1）の評価である．なお，Web サイト脅威度推定機能へのもう 1 つの要件に，（要件 2）望ましくない認証結果の検証がある．ただし，この要件のための機能である認証結果精査機能の評価には，認証結果が必要なものの，Web サイト脅威度推定機能は，実証実験までに至っていないことから，この認証結果が得られていない．このため，同機能の評価は，今後の課題とする．

(1)　情報取得可能性

3 章で述べたとおり，Web サイト脅威度推定機能は，Web サイトの脅威度を推定するために，URL 等の静的に取得できる情報に加えて，WHOIS 情報や DNS 情報等の外部サイトを参照して取得する必要のある情報を利用する．このため，Web サイトの脅威度を推定するに先立って，外部情報を取得できることが Web サイト脅威度推定機能に要求される．そこで，外部情報の取得可能性を検証する．

(2)　脅威度推定精度

Web サイト脅威度推定機能は，複数の学習器を利用して，アクセス先サイトの脅威度を推定する．良性データ/悪性データを用いてこの推定精度を検証する．

(3)　追加認証数抑制度

従来の AED において解決するのが望ましい課題として，（課題 1）に挙げたように，追加認証が頻発し得るというものがある．Web サイト脅威度推定機能は本課題への対応を 1 つの要件として設計したものである．そこで，従来の AED では発生していた追加認証を Web サイト脅威度推定機能によってどれだけ抑制できるか検証する．

4.3 情報取得可能性

良性サイトのサンプルとして，オープン Web ディレクトリである DMOZ [20]から収集したサイト群，悪性サイトのサンプルとして hpHosts[21], spamhaus[22], Malware Domain List[23], および aguse[24]から収集したサイト群を利用した．この際，良性/悪性サイト情報それぞれ 50,000 件ずつ，合計 100,000 件取得した．これらの情報を以降の節での評価に利用する．また，実験中，DNS 情報や WHOIS 情報を取得できない場面が特に悪性サイトに見られた．これは，Web サイト群を収集した時点，あるいは収集してから当該サイトの情報を取得するまでの間にドメインが失効したことが原因ではないかと考えられる．特に，悪性サイトの生存期間は，良性サイトに比べて短い傾向があり，短いものは数日でドメインが失効するものもあると報告されている [7][8][9][10]．ただし，実際に運用する際には，ドメインが失効したサイトに対しては，接続が成立し得ない．このため，AED の防御対象である悪性サイトへのアクセスに伴う被害は生じ得ないことから，問題ないと考えられる．

また，情報の取得に成功したものは，DB に保存し，再利用できる形としている．実際に取得したデータのうち，一部抜粋したものを図 3 に示す．

図 3 から，WHOIS 情報（WHOIS），DNS 情報（DNS），地理情報（geo_info），および Alexa ランク（Alexa）の取得に成功していることが分かる．

以上のように，外部情報取得機能は，脅威度推定のために利用する情報を取得し，かつ DB に保存できていることを確認した．


```
{
 "domain": "hitachi.co.jp",
 "url": "http://www.hitachi.co.jp/",
 "Alexa": {
  "POPULARITY": 10741,
  "REACH": 11319
 },
 "WHOIS": {
  "updated_date": [
   "2016-04-01T01:03:42"
  ],
  "nameservers": [
   "ns1.hitachi.co.jp",
   "ns2.hitachi.co.jp",
   "ns3.hitachi.co.jp"
  ],
                                                    ...
},
 "geo_info": {
  "country_name": "Japan",
  "asn": "2526"
                                                    ...
 },
 "DNS": {
  "A": [
   "133.145.228.249"
  ],
"rev": [
    "249.228.145.133.in-addr.arpa."
  ],
  "MX": [
   "hitiij-kan.hitachi.co.jp.",
   "hitpro.hitachi.co.jp.",
   "hitiij.hitachi.co.jp."
                                                    ...
 }
}
```


図 3　外部情報のサンプル（一部抜粋）

4.4 脅威度推定精度

本評価では，4.3 節で収集した良性サンプル 50,000 件と悪性サンプル 50,000 件を用いて，提案手法の脅威度推定精度を検証する．まず，URL 文字列のみを用いた場合，WHOIS 情報のみを用いた場合，および DNS 情報+Alexa ランク+地理情報を用いた場合の 3 パターンにおいて，各種アルゴリズム（線形 SVM，ロジスティック回帰，決定木，K 近傍法，ランダムフォレスト，3 層ニューラルネット，およびアダブースト）を用いて評価を実施した．その後，全情報をまとめて 1 つの推定器に利用した場合と各パターンで最も高精度であったアルゴリズムのものを組み合わせた場合の提案手法で比較評価を行った．各アルゴリズムでは，対象 Web サイトの脅威度を 0（良性寄り）～100（悪性寄り）の連続値で算出するが，この際，悪性サイト/良性サイトの分類閾値には，50 を用いた（脅威度が 50 未満の場合：良性と判断，脅威度が 50 以上の場合：悪性と判断）．なお，提案手法において，各推定器の重みには，同データセットに対する各推定器の正解率の比率を利用した．さらに，提案手法に関しては，閾値をどの値に設定すれば良いか，過検知率（False Positive Rate，以降 FPR），見逃し率（False Negative Rate，以降 FNR），および正解率の観点から評価した．なお，これらの評価には，10-分割交差検証を用いた．

まず，各パターンにおける脅威度推定精度の測定結果を表 3 に示す．表 3 から，URL 文字列のみ，WHOIS 情報のみを用いた場合はランダムフォレスト，DNS 情報+Alexa ランク+地理情報を用いた場合は K 近傍法がそれぞれ 68.92%，83.07%，および 88.79%と最高精度であることが確認できる．さらに，上述のアルゴリズムを用いた 3 つの推定器を組み合わせて利用したところ，各推定器単体のいずれ（68.92%，83.07%，および 88.79%）よりも高く，かつ全ての情報を単純にまとめて利用した場合（91.26%）よりも高い 92.74%の精度で Web サイトを分類できた．また，図 4 は，各パターンでの最高精度を出した分類器の Receiver Operating Characteristic（以降，ROC）曲線である．ROC 曲線において，曲線下の面積を Area Under the Curve （以降，AUC）と呼び，AUC が 1 に近いほど，識別性能が高いことを示す．図 4 から，提案手法の AUC が 0.976 と正解率と同様に他のいずれのパターンにおける値（0.754，0.903，0.949，および 0.966）よりも高いことが分かる．

以上のように，提案手法が各推定器単体の場合や全情報をまとめて 1 つの推定器で利用した場合のいずれよりも高精度で Web サイトの脅威を推定できた．また，複数の文献において，URL 文字列のみで推定することの問題点が示唆されている（精度の確保が難しい[4]，短縮 URL を誤判定してしまう[5]等）が，本実験においても，URL 文字列のみを使ったものは，最高でも 68.92%と他に比べて精度が低く，各文献での示唆内容を裏付けるものとなった．

次に，提案手法において，閾値を 0～100 の間で変動さ

表 3　各パターンにおける脅威度推定精度の測定結果（%）

	URL	WHOIS	DNS	全て利用	提案手法
線形SVM	66.11	81.72	74.32	86.72	
ロジスティック回帰	66.11	81.73	74.38	86.76	
決定木	68.81	82.06	84.30	88.67	
K近傍法	67.75	80.75	88.79	83.80	
ランダムフォレスト	68.92	83.07	85.53	91.26	
ニューラルネット	67.07	82.27	80.00	88.72	
アダブースト	67.17	81.98	81.17	89.64	
最高精度	68.92	83.07	88.79	91.26	92.74

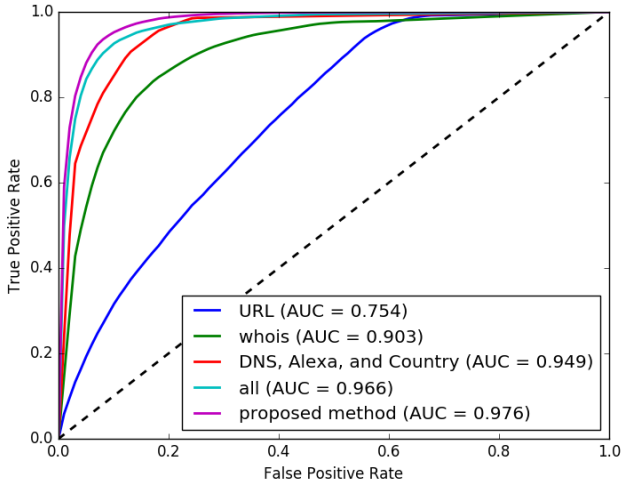

図 4　各パターンにおける ROC 曲線

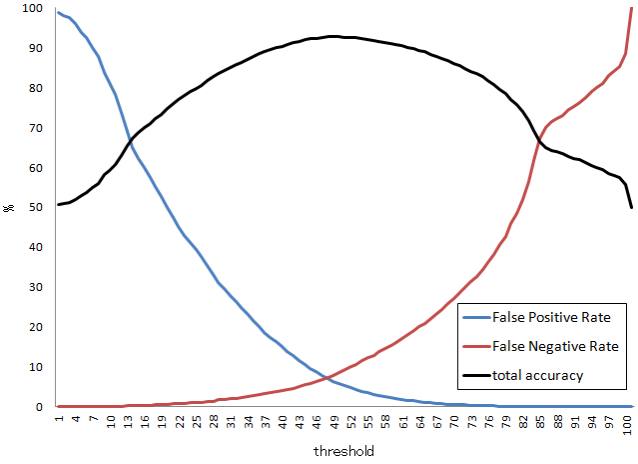

図 5　閾値毎の FPR，FNR，および正解率（%）

せ，それぞれの値における FPR，FNR，および正解率を算出した．この結果を図 5 に示す．

図 5 から分かるように，閾値を 48 に設定することで，92.82%と最も高い正解率を得られている．また，閾値を下げると良性なサイトであっても不審サイトとして過検知してしまう可能性（FPR）が高まり，反対に閾値を上げると悪性サイトを見逃してしまう可能性（FNR）が高まる．ここで，悪性サイトを見逃してしまうと，同サイトへのアクセスが発生して被害が生じうることや良性サイトを不審サイトとして過検知してしまった場合でも，AED であれば追加認証さえ突破すればアクセスは続行でき，業務効率への影響は最小限に抑制できることから，FPR の増大はある程

表 4　追加認証抑制数

	合計	追加認証抑制数	
		閾値:48	閾値:37
Case 1	68,653	58,349 (84.99 %)	49,683 (72.37 %)
Case 2	57,842	40,976 (70.84 %)	30,298 (52.38 %)

度許容できるとともに，FNR を抑制することが望ましい．最高精度を出せる閾値 48 の場合は，FNR が 7.36%だが，これを半分以下に抑えようとした場合, 閾値を 37 にすることにより, 正解率を 89.96%に留めつつ達成できる(3.518%)．

以上の評価結果から，複数の推定器を用いることによって，単体の場合よりも精度が出せること，正解率の視点からは閾値を 48 にすれば良いことが分かった．

4.5 追加認証抑制度

本評価では，我々の組織内で従来の AED の実証実験を実施した際，グレーリストに属するサイトへのアクセスであるとして追加認証画面を出す元となった URL を用いて，Web サイト脅威度推定機能の追加認証抑制度を評価した．ここで，実証実験中に真に悪性なサイトへのアクセスは確認されなかったことから，同実験中の追加認証要求は，本来は不要なものである．この内，何%を抑制できるか実験することにより，実証実験の際に本機能があれば，不要な追加認証がどの程度抑制できていたかを仮想的に評価する．また，実証実験は，ホワイト/ブラックリストにないアクセス先全てをグレーリストとみなした場合（Case 1）とインテリジェンスを用いてグレーリストを作成した場合（Case 2）の 2 パターン実施している．本評価では，その両方の場合に対し追加認証抑制度を検証した．

訓練データセットには，4.2 節で収集した 100,000 件を，良性サイト/悪性サイトの閾値には，4.3 節の評価において総合的に最も高い正解率を出した 48 と FNR を半分以下に抑制できる 37 の 2 つを利用した．また，上記のデータセットを用いて作成した推定器に Case 1，Case 2 でそれぞれ 68,653 件，57,842 件の評価用データを投入し，本検証を行った．評価結果を表 4 に示す．

実験では，閾値を正解率優先にした場合で 84.99%（Case 1）・70.84%（Case 2），安全性優先にした場合 72.37%（Case 1）・52.38%（Case 2）の追加認証を抑制できた．

Case 1 のホワイト/ブラックリストにないアクセス先全てをグレーリストとみなした場合は，Case 2 のインテリジェンスを用いてグレーリストを作成した場合よりも多くの追加認証を抑制できている．これは，明らかに良性なサイトは，未知のアクセス先を全てグレーリストとみなす関係上，前者には含まれ得るのに対し，インテリジェンスをベースにしている後者には，そのようなサイトの含まれる可能性が相対的に低いことが原因であると考えられる. また，この結果から，Web サイト脅威度推定器の妥当性を間接的にではあるが証明できた．

また，閾値を下げ，安全側に倒した場合であっても 50～70%程度の不要な追加認証は抑制できた．ただし，閾値を下げると追加認証数は増加してしまうことから，FNR と追加認証抑制度の間にはトレードオフの関係があり，閾値を上下させることにより，どちらを重視するかを変更できる．追加認証抑制度を重視する場合，訓練データを最高精度で分類できた閾値 48 を採用することにより, 不要な追加認証を 84.99%抑制できる．

以上のように，Web サイト脅威度推定機能に妥当性があること, および, 同機構によって不要な追加認証を最大 85%程度抑制できることを確認した．

5. 関連研究

本章では，提案手法と同様に，Web サイトの性質を判別する研究について述べる．また，同研究は，以下の３種類に大別できる．

(1)　Web サイトへのアクセスが不要なもの

文献[4]は，既知の悪性 URL 群と近い性質を持った URL を未知の URL 群から抽出し，Bayesian sets と呼ばれる類似要素探索アルゴリズムを用いてブラックリストを構築することを目指したものである．文献[5]は，URL 文字列ベースで決定木によって良性/悪性サイト分類を行っている．文献[6]は, URL のみからフィッシングサイトか否かを判定するものである．Page Rank やドメインにフィッシング特有の文字列が出現しているか否かを特徴量とし，ロジスティック回帰により判定を行っている. 各研究の利点として, Web サイトへアクセスしないため，比較的判定が高速なことや攻撃者のアクセスログ等を取ることによる被解析検知を逃れられることが挙げられる．一方で，文字列のみを用いる場合, 後述する研究のような Web サイトにアクセスして情報を取得する場合に比べると取得できる情報に限りがあり, 精度の面ではやや劣る傾向が見られる．

(2)　非悪性サイトへのアクセスが必要なもの

PREDETOR[3]は，ドメインの文字列や登録情報を用いて悪性か否かを分類するものである．文献中で，悪用されるドメインの登録にはバースト性があることや失効したドメインが即時再取得された場合は，攻撃者による取得である可能性が比較的高いこと等が述べられており，悪性ドメインの分類に寄与することが実証されている．文献[7]は，URL 文字列やホスト情報を用いた悪性サイトの推測器を提案している文献であり，フィッシングに関連する URL は，そうでない URL に比べて URL が長いこと，ドメイン名が長いこと，およびドメインの生存期間が短いこと等が検証結果として示されている．文献[10]は，クライアント型ハニーポットでの URL の巡回を最適化するために, 巡回候補 URL の悪性度を算出し, その高い順に巡回するものである．この悪性度を算出するために，SVM を利用し，特徴量には whois 情報や FQDN 文字列の特徴を採用している．文献[12]は, 悪性 Web サイトが属する IP アドレスブロック

とドメイン登録に用いたレジストラに着目し，両情報が既知の悪性ドメインのものと類似している場合，信頼性が低いドメインであるとして，ブラックリストに追加する手法を提案している．EXPOSURE[14]は，主に DNS 情報を用いて決定木で悪性ドメインを見つけるものである．これらの研究は，Web アクセスが不要な研究よりも推測に利用できる情報が多いため，精度が比較的高い傾向が見られる．一方の欠点としては，Web アクセスが不要な(1)に比べると，推測に要する時間が比較的長いことや対象の Web サイトにアクセスする(3)に比べると取得できる情報に限りがあり，精度が劣る傾向が見られる点が挙げられる．

(3)　悪性サイトへのアクセスが必要なもの

EvilSeed[9]は，既知の悪性サイト情報を基に，クライアント型ハニーポットが利用する効率的な巡回クエリを生成し，悪性サイトを収集するものである．既知の悪性サイトリストを基（Seed）にして，リンクや DNS 情報等が類似しているページを検索するクエリを生成する．そのクエリを基に巡回したページの性質を Oracle と呼ばれる Google Safe Browsing 等からなるコンポーネットを用いて判定し，悪性と判断されたものは，再度検索クエリ作成の Seed とすることで効率的に悪性サイトを巡回することが可能となる．本研究のように，実際に悪性サイトへアクセスするものは，比較的時間を要する・クローキングへの対策が必要である等の懸念があるものの，精度の面で優れている傾向にある．

提案手法は，複数の推定器を組み合わせるため，上述した手法等をその構成要素の 1 つとして活用することが可能である．また，今回は，(1)と(2)に分類される手法のみを構成要素として用いたが，(3)にあたる手法を組み込むことによって，更なる分類精度向上が期待できる．

6. おわりに

本稿では，AED の適用に伴う業務への悪影響を削減するため，Web サイト脅威度推定機能の検討，開発，および評価を行った．評価によって，サイトの性質を最大 92.82%の精度で推定できること，追加認証数を最大 84.99%抑制できることが確認された．これにより，AED の課題の 1 つであった追加認証が頻発し得る点を抑制できたといえる．

今後の課題には，もう 1 つの課題である望ましくない認証結果が混入することによる推定精度低下抑制を目的とした認証結果精査機能の実現と評価がある．また，Web サイトの脅威度推定方式をより改善し，推定精度を維持しつつ FNR を抑制することもある．

参考文献

[1] 仲小路博史，藤井康広，磯部義明，重本倫宏，鬼頭哲郎，川口信隆，林 直樹，下間直樹，菊池浩明：人間行動を用いた自律進化型防御システムの提案，2016 年暗号と情報セキュリティシンポジウム（SCIS2016），pp.1－8 (2016).

[2] Nakakoji, H., Fujii, Y., Isobe, Y., Shigemoto, T., Kito, T., Hayashi, N. Kawaguchi, N., Shimotsuma, N., and Kikuchi H.: Proposal and Evaluation of Cyber Defense System Using Blacklist Refined Based on Authentication Results, The 19th International Conference on Network-Based Information Systems (NBiS2016), pp. 135－139 (2016).

[3] Shuang, H., Alex, K., Brad, M., Vern, P., and Nick, S.: PREDATOR: Proactive Recognition and Elimination of Domain Abuse at Time-Of-Registration, Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16), pp.1568－1579 (2016).

[4] 孫博，秋山満昭，八木毅，森達哉：既知の悪性 URL 群と類似した特徴を持つ URL の検索，コンピュータセキュリティシンポジウム 2014（CSS2014）論文集，pp.1－8 (2014).

[5] Michael, D., Greg, H., Gilad, G., Aravind, A., and Prabaharan, P.: A Lexical Approach for Classifying Malicious URLs, 2015 International Conference on High Performance Computing (HPCS 2015), pp.195－202 (2015).

[6] Sujata, G., Niels, P., Monica, C., and Aviel D. R.: A framework for detection and measurement of phishing attacks, Proceedings of the 2007 ACM workshop on Recurring malcode, pp.1－8 (2007).

[7] Ma, J., Saul, L. K., Savage, S., and Voelker, G. M.: Beyond blacklists: learning to detect malicious Web sites from suspicious URLs, Proc. of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2009), pp.1245－1254 (2009).

[8] Mark, F., Christian, K., and Vern, P.: On the potential of proactive domain blacklisting, Proceedings of the 3rd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more (LEET '10), pp.1－8 (2010).

[9] Invernizzi, L., Benvenuti, S., Cova, M., Comparetti, P. M., Kruegel, C. and Vigna, G.: EvilSeed: A Guided Approach to Finding Malicious Web Pages, Proc. of IEEE Symposium on Security and Privacy, pp.428－442 (2012).

[10] 千葉大紀，森達哉，後藤滋樹：悪性 Web サイト探索のための優先巡回順序の選定法，コンピュータセキュリティシンポジウム 2012（CSS2012）論文集，pp.805－812 (2012).

[11] 福島祥郎，堀良彰，櫻井幸一：ドメイン情報に着目した悪性 Web サイトの活動傾向調査と関連性分析，コンピュータセキュリティシンポジウム 2010（CSS2010）論文集 (2012).

[12] 福島祥郎，堀良彰，櫻井幸一：悪性 Web サイト間の関連性に着目した信頼性評価によるブラックリスト方式の検討，情報処理学会研究報告，Vol-CSEC-52 No.38，pp.1－8 (2011).

[13] Ricardo, V. S., and Jose, C. B.: Identifying Botnets Using Anomaly Detection Techniques Applied to DNS Traffic, 5th IEEE Consumer Communications and Networking Conference, pp.476－481 (2008).

[14] Leyla, B., Engin, K., Christopher, K., and Marco, B.: EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis, 18th Annual Network & Distributed System Security Symposium (NDSS '11) (2011).

[15] scikit-learn: machine learning in Python, available from <http://scikit-learn.org/stable/> (2017-03-09 accessed).

[16] joepie91: GitHub - joepie91/python-whois: A python module for retrieving and parsing WHOIS data, available from <https://github.com/joepie91/python-whois> (2017-02-23 accessed).

[17] nominum: dnspython home page, available from <http://www.dnspython.org/> (2017-02-23 accessed).

[18] Amazon Web Services, Inc.: AWS | Alexa Web Information Service - Traffic Metrics for any Website , <https://aws.amazon.com/jp/awis/> (2017-02-23 accessed).

[19] MAXMIND: https://www.maxmind.com/ja/home, available from <https://aws.amazon.com/jp/awis/> (2017-02-23 accessed).

[20] dmoz: DMOZ - The Directory of the Web, available from <https://www.dmoz.org/> (2017-03-03 accessed).

[21] hpHosts: hpHosts Online - Simple, Searchable & FREE!, available from <http://www.hosts-file.net/> (2017-03-06 accessed).

[22] SPAMHAUS: The Spamhaus Project, available from <http://www.spamhaus.org/> (2017-03-06 accessed).

[23] Malware Domain List: MDL, available from <http://www.malwaredomainlist.com/> (2017-03-06 accessed).

[24] aguse：aguse.jp：ウェブ調査，入手先 ＜https://www.aguse.jp/> (2017-03-06 アクセス).