セキュリティインシデント解析の為の Web Mining システム

山口 崇志† 谷出 広大† 三須 剛史‡ 岩崎 信也‡
岸本 頼紀† 花田 真樹† 布広 永示†

†東京情報大学 総合情報学科
265-8501 千葉県千葉市若葉区御成台 4-1
tyamagu@rsch.tuis.ac.jp

‡東京情報大学大学院総合情報学研究科
265-8501 千葉県千葉市若葉区御成台 4-1
tyamagu@rsch.tuis.ac.jp

あらまし 標的型攻撃のような近年の多様で高度化したサイバー攻撃においては，従来のような通信ログを基とした解析により攻撃の検知や攻撃者の特定，さらにはその動機を知ることが困難で，セキュリティインシデントに直面した際の迅速且つ適切な対策が難しい．我々は通信ログ以外のデータも用いて解析し，セキュリティインシデントの早期発見や攻撃者とその動機を推定する研究を進めている．本稿では，我々の構築した自然言語処理の技術を応用してWeb上の文章からイベントやキーワードの抽出を行う基盤システムについて解説する．

Web Mining System for Security Incidents Analysis

Takashi Yamaguchi† Koudai Tanide† Takeshi Misu‡ Shinya Iwasaki‡
Yorinori Kishimoto† Masaki Hanada† Eiji Nunohiro†

†Tokyo University of Information Sciences, Department of Informatics
4-1 Onaridai, Wakaba-ku, Chiba, Chiba Prefecture 265-8501, JAPAN
tyamagu@rsch.tuis.ac.jp
‡Tokyo University of Information Sciences, Graduate School of Informatics
4-1 Onaridai, Wakaba-ku, Chiba, Chiba Prefecture 265-8501, JAPAN
tyamagu@rsch.tuis.ac.jp

Abstract In the case of targeted thread, the attacker uses the various attack methods that include the social engineering to the particular organization with the definite purpose. Therefore, the detection and the security measures are difficult by the past detection technologies based on network log analysis in these complicated cyber incident. In this research, we develop a web-mining system in order to catch the sign of cyber attacking from the documents on Web. In this paper, we proposed event extraction method based on natural language analysis for the extraction of events on cyber-attacks.

1　はじめに

近年日本国内の企業では，サイバー攻撃への遭遇率が増加しており，特に標的型攻撃と思われる事例が多く報告されている[1]．標的型攻撃とは，特定の企業や公共機関に対して，情報の窃取等の明確な目的を持った一連の攻撃である．個々の攻撃手法としては，特に電子メールを媒介とし取引先からの連絡に偽装する等，ソーシャルエンジニアリングの悪用によるものが多く報告されている．電子メールを利用した手口以外にも，ランサムウエアや水飲み場型攻撃のように人間の心理を突いた新しい攻撃手法が増加をみせている．これらの攻撃手法は，人間の心理的な隙をついて人為ミスを誘うことから，一般的なコンピュータセキュリティソフトウェアによる未然の検査やフィルタリングが難しく，情報の窃取等を目的とする標的型攻撃においては初期のバックドアとなる不正プログラムの投入に用いられることが多い．

一方で，サイバー攻撃においてソーシャルエンジニアリングが悪用される場合，必ずその攻撃の目的を達成する為に，ソーシャルエンジニアリングに基づく対象の絞り込みや詐称等の方策が取られており，これらに注目して分析する事により，標的型攻撃の初動を検知するとともに，攻撃者の目的や場合によっては攻撃者そのものを推定することが期待できる．従来，これらの分析は，分析に関わったセキュリティ技術者の経験や経済，国際情勢等に関する幅広い知識に依存することが多く，これら関連する多様な情報を関連付け分析をサポートする仕組みが必要である．

本研究では，通信ログ以外の多様なデータも用いて解析し，セキュリティインシデントの早期発見や攻撃者とその動機を推定する研究を進めており，Web（World Wide Web）を中心としたデータ収集と自然言語処理の技術を応用した分析と可視化をするシステムを構築した．本稿では構築したセキュリティインシデント解析システム（SIAS: Security Incident Analysis System）の解説をする．

2　提案解析システムの構成

本研究では Web を中心にデータを収集し，収集したデータを自然言語処理の技術を応用してキーワードやイベントの抽出を行う為に，Web から情報を収集する Web クローラと自然言語処理を行うモジュール群および可視化ツールを連携する解析システムを構築した．最初に解析システムの概要を示し，中核となる Web クローラ，自然言語処理モジュールの構成や関連技術について説明する．

提案する解析システムはセキュリティインシデントに関するデータマイニングのプロセスを補佐する為のシステムで，一般的なビジネス・インテリジェンス（BI）システムの構成に近い．一般的なデータマイニングのプロセスと BI システムの構成を図 1 に示す[2][3]．

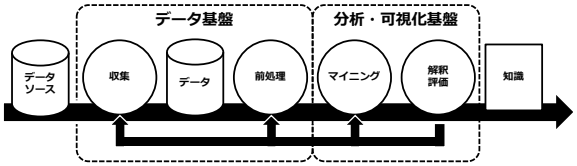

図 1. 一般的なデータマイニングのプロセスと BI システムの構成

通常データマイニングではデータソースから得た未加工のデータから必要な情報を抽出し，データ形式の変換や符号化，インデックス化等の加工処理をした後，分析用データベースに記録される．その後，統計や機械学習等を用いた解析処理を行い，必要に応じて可視化をした後，人間の手によって評価や解釈が行われる．データマイニングのプロセスを補佐する BI システムでは，主に前半のデータ加工の処理をデータ基盤，後半を分析・可視化基盤として一般化し構築されることが多い[3]．

本研究では，データソースが Web 上の文章で且つ情報抽出に自然言語処理を応用することから，図 1 のマイニングプロセスに多様なモジュール群とマイニング時の試行錯誤が必要になる．また，自然言語処理を大量のデータへ適

用する場合，文章や単語に対する特徴量は永続的に保持しておくことが望ましい．したがって，データベースを中心とした構成とし，分析結果のデータ構造を保持可能な抽象的なデータ型と対応するデータ形式を定義することで，前処理，分析，可視化等，各モジュールの祖結合化を行い，可視化等外部ツールを利用し易い方式をとっている．また，すべての操作はWeb APIを通じて行い，データベースからの出力結果はXML, JSON, CSV 形式で出力可能である．

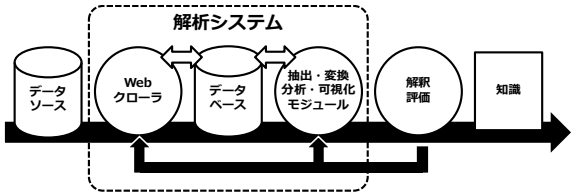

図 2. データマイニングのプロセスと提案する解析システムの構成

2.1 DOM ベース Web クローラ

一般的な Web クローラでは複数のソフトウェアロボットを用いて Web サイトを巡回し，Web サイト上のファイル群を収集する．オープンソース等で開発が進められているWebクローラは，簡易的な HTTP 通信により Responce Body を保存しており，現在主流である Java Script によりDOM(Document Object Model)[4]を構築している Web ページでは十分に情報を収集することができない．したがって，本研究ではJava Scriptを実行しDOMの変化を収集するDOM ベース Web クローラを開発した．

開発した Web クローラは，一般的な Web ブラウザ同様，HTML を DOM ツリーとして扱い，Java Script によるDOMツリーの変化を監視し時間と共に記録する．一般的なWebブラウザではHTMLからDOMツリーを構築し，付随する Java Script での操作を実行した上で，DOMツリーからHTMLへ再変換し描画しているが，本クローラも同様に DOM ツリーに変化のあった箇所を HTML へ再変換し記録している．記録する情報を表 1 に示す．初期状態のDOMツリーが構築された時間をセッション開始時間とする．また，これ以外に従来のクローラと同様，HTTP の Request と Response の Header および body を時間と共に記録している．

表 1. Web から取得し記録する情報

属性名	*データ型*
URI	文字列
セッション開始時間	整数値
DOM 変更時間	整数値
HTML コンテンツ	文字列

DOM ベース Web クローラは，DOM からイベントを取得できる為，Java Script によって記述されたイベント処理を能動的に呼び出すことも可能である．一方で，悪意のあるコードも実行される為，これらに対する対策と併用するのが望ましい．

Web クローラによって収集されたデータはそのままデータベースへ保持される．一般的なクローラと同様に URL リンクの抽出は次節 2.2 に示す抽出・変換・分析モジュールの一つとしてデータが記録される毎に実行される．抽出した URL 郡はデータベースへ一時的に記述され，Web クローラのボット郡はこれらのデータを参照し巡回，収集する．

2.2 抽出・変換・分析モジュール

抽出・変換・分析モジュールは，全てデータベースよりデータを受け取り，対応する処理を行った後，その結果をデータベース上に記録する形に一般化できる．出力する結果は抽象的なデータ構造のみ定義されており，データ構造の型が一致する限り，多様なモジュール郡を連携して処理可能である．

表 2 および表 3 に代表的なモジュールと入出力のデータ構造とデータ型の例を示す．ここで，データ構造の型はリストやグラフ等，データ構造を示す情報で，データ型やデータ構造型で示される型に一致する値を格納するコンテナをデ

ータ構造に従って順序付ける一種のインデックスである. データ構造をある種のインデックスとして定義することで, 特定のデータ構造に対し適切な順序でアクセスし繰り返し行う処理を容易に記述できる. なお, データ型は分析の結果得られる特徴量を保持する為, XML のように自由に拡張し属性値を追加できる必要があり, データベース管理システムは一般的なリレーショナル・データベースでの実装よりMongoDB[5]や OrientDB[6]等に代表されるドキュメントデータベースあるいは複雑なインデックス化が可能なデータベースが望ましい.

表 2. モジュール群と入出力のデータ構造の例

モジュール名	*入力型*	*出力型*
URL リンク抽出	単一<文章>	集合<文字列>
トークン化	単一<文字列>	集合<トークン>
キーワード抽出	単一<文章 ID>, 単一<文章集合>	集合<トークン>
イベント抽出	単一<文字列>	単一<イベント>
URL リンク構造抽出	リスト<文章>	単一<グラフ>

※入出力の記述書式:データ構造型<データ型>

表 3. データ型の例

モジュール名	*属性値名*	*データ型*
文字列	文字列	単一<文字列>
トークン	文字列	単一<文字列>
文章	文章 ID	単一<文字列>
	文章	単一<文字列>
文章集合	文章集合 ID	単一<文字列>
	文章集合	集合<文章>
グラフ	グラフ ID	単一<文字列>
	ノード集合	集合<ノード>
	エッジ集合	集合<エッジ>
ノード	ノード ID	単一<文字列>
	文字列	文字列
エッジ	開点ノード ID	単一<文字列>
	終点ノード ID	単一<文字列>
イベント	対象の文章 ID	単一<文字列>,
	属性キー	集合<文字列>
	属性値	KV< 単一<文字列>, 単一<*> >

※データ型の記述書式:データ構造型<データ型>

※Key-Value 型は Key と対応する Value を持つ型で, それぞれの型を KV<Key の型,Value の型>と記述する

具体的な実装例について URL リンク抽出と URL リンク構造抽出を例に説明する. URL リンク抽出では入力として Web 文章を受け取る. Web 文書は URL を“文章 ID”とし, 対応する HTML 文章を属性値“文章”として定義する. DOM 型 Web クローラではある URL に対し, DOM の更新毎に複数の HTML データが保持されるが, これらを全てマージしたものを“文章”の属性値としている. 実際の URL 抽出処理は, 受け取った“文章”の値である文字列からパターンマッチングにより URL リンクとして抽出しリスト型文字列で格納する.

URL リンク構造抽出では, HTML 内の URL のリンク構造を頻度を重みとする重み付き有効グラフとして抽出する. 文章のリストを入力として受け取り, 各文書に対して URL リンク抽出を実行する. その結果得られた URL のリストと“文章 ID”である URL から URL を“ノード ID”とするノードのリストと始点と終点のノード ID をキーとするエッジのリストを生成すると共に, ノード, エッジ共に属性値として“頻度”を追加し出現頻度をカウントする.

記録されたグラフ型データはグラフ ID によりアクセスし XML や JSON, CSV 形式で取り出すことができる. これにより, グラフの可視化・分析ツール[7]や D3.js[8]等によるグラフ可視化 Web アプリケーションとの連携が容易に可能である.

2.3 自然言語処理と Web マイニング

データマイニングの応用として, 主に自然言語からの特徴抽出に注目した分野をテキストマイニング[9]と呼び, 得に Web 上の文章や HTML 等から特徴を抽出することを Web マイニングと呼ぶ[10][11][12]. 近年の自然言語処理の研究は, 意味を理解することを目標とした自然言語理解 [13]と, 現在幅広く応用されている統計的な解析を主とした自然言語処理に分けることができる.

一般的に自然言語理解に関する研究は発展途上であり, 近年, 実応用で成果をあげているのは, Web の検索システムに代表されるような統計処理を主体とした特徴抽出手法 [14][15][16] によるものである. 本研究では,

文章を短いトークンに分割し特徴量を導き出すN-Gram[14]ベースの解析モジュールに加え，キーワード抽出による特徴量抽出手法であるTF-IDF[15]やその発展で確率分を考慮して特徴量の次元縮約を行うLDA（Latent Dirichlet Allocation）[16]等の代表的な特徴量抽出モジュールを実装した．また，純粋なN-Gramでは言語非依存な処理が可能な反面，意味のないトークンが生成されることから，形態素解析によるトークン化処理も扱えるようにしている．

形態素解析は基本的に自然言語の文法に依存することから，単一のアルゴリズムで多言語対応するのが難しい．したがって，本システムでは，HTMLの文字コードより言語を識別し対応する言語の形態素解析モジュールを用いる．特に日本語においては名詞が実際の意味を持つ単位より細かく分割される傾向があることから，構文解析を応用しトークンを生成する．代表的な構文解析手法に確率文脈自由文法（SCFG: Stochastic context-free grammar）[17]が挙げられるが，多大な計算量を必要とする為，連続する名詞句を単純に結合するトークン結合モジュールも実装した．単純にキーワード抽出を行うだけの場合，簡易的なトークン結合処理のみで十分である．

2.4　固有表現とその関係の抽出

2.3で示したトークン化処理を施すことで文章は単語の配列に分割され，トークン毎に頻度や共起頻度を測ることで文章の特徴として扱うことが可能である．一方で，ここで得られるトークンは単語と品詞の情報のみな為，イベント抽出等を考慮した場合，トークン毎に組織名や人名，地名，日付や時間等に割り当てる必要がある．

このような処理を固有表現抽出と呼び，一般的には分類問題として考えることが可能で，Support Vector Machine[18]やConditional Random Field）[19]等機械学習を用いた手法で比較的良い結果が報告されている[20]．しかしながら，特に日付や時間の抽出に関しては，多様なパターンの作成コストが高いものの，人の手によって作成したパターンに基づく抽出の方が高い精度が得られる可能性が高い．

本研究では特に日付に関しては人の手によって作成したパターンによる抽出を行い，組織名，人名，地名については機械学習によって作成した分類器を実装した．

3　インシデント解析への適応

2章で解説した解析システムは汎用的なWebマイニングシステムであることから，セキュリティインシデン解析に応用できるよう適応を行った．適応の主軸は多言語への対応で，前述のように形態素解析によるトークン化や固有表現抽出に関しては言語による依存性が非常に高い反面，マルウェアの配布元の状況を考慮すると，英語圏だけでなく，中国語やロシア語圏の文章を解析する必要がある．そこで，2.3節でも述べたように文字コードにより自然言語解析モジュールを使い分ける手法をとっている．しかしながら，形態素解析の手法については，多くの場合各言語圏において開発が進められており，特に中国語に関する形態素解析機に関しては今後の課題としている．

次に各種ソーシャルメディア等のWebサービス対応が挙げられる．HTML文章は，本来中心となる文章コンテンツ以外に広告やUIに関連する文字列が含まれており，精度低下の要因となることが多い．したがって，特に利用頻度の高いサービスに関してはサイト毎のパターンマッチングによる情報抽出やAPI経由でのアクセスを用いた方が精度の向上が期待できる．したがって本研究では，Webクローリング時にURLのホスト名により処理を分岐し，特定のWebサイト専用の情報抽出やAPIアクセスモジュールを用いている．

4　おわりに

本研究では，通信ログ以外の多様なデータも用いて解析し，セキュリティインシデントの早期発見や攻撃者とその動機を推定する為，Webからデータ収集を行い自然言語処理の技術を応用して解析するシステムを構築した．

現在データ収集を進めており，今後は本解析システムを用いた分析結果について実証のもと評価を行う．具体的には 2 つのアプローチを予定しており，ソーシャルメディアやチャットログを中心とする Hacktivist の投稿からのキーワード抽出と，セキュリティに関するニュースやソーシャルメディアにおけるセキュリティ関連のキーワード抽出とイベント抽出を行う．特にセキュリティに関しては一般のユーザ，セキュリティ専門家，攻撃者の各グループにおいて，用いている語彙が異なっていると考えられ，これらを解析する為の基盤となるキーワードの抽出は重要である．また，イベント抽出では，セキュリティに関するイベントを機械敵に要約するシステムへの応用が期待される．

参考文献

[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] U.M. Fayyad, A. Wierse, G.G. Grinstein, Information visualization in data mining and knowledge discovery, Morgan Kaufmann, 2002

[3] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, Elsevier, 2011

[4] W3C, Document Object Model (DOM) Specifications, http://www.w3.org/DOM/DOMTR

[5] MongoDB Inc. , MongoDB, https://www.mongodb.org/

[6] Orient Technologies LTD , Orient DB, http://orientdb.com/about-us/

[7] Shannon, Paul, et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome research 13.11 (2003): 2498-2504.

[8] Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer. "D^3 data-driven documents." Visualization and Computer Graphics, IEEE Transactions on 17.12 (2011): 2301-2309.

[9] Feldman, Ronen, and James Sanger. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, 2007.

[10] Kosala, Raymond, and Hendrik Blockeel. "Web mining research: A survey." ACM Sigkdd Explorations Newsletter 2.1 (2000): 1-15.

[11] Cooley, Robert, Bamshad Mobasher, and Jaideep Srivastava. "Web mining: Information and pattern discovery on the world wide web." Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on. IEEE, 1997.

[12] 奥村 学, ソーシャルメディアを対象としたテキストマイニング, 電子情報通信学会 基礎・境界ソサイエティ Fundamentals Review 6(4), 285-293, 2013

[13] Jurafsky, Dan, and James H. Martin. Speech and language processing. Pearson, 2014.

[14] Shannon, Claude Elwood. "A mathematical theory of communication." ACM SIGMOBILE Mobile Computing and Communications Review 5.1 (2001): 3-55.

[15] Salton, Gerard, and Michael J. McGill. "Introduction to modern information

retrieval." (1986).
[16] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.
[17] Sakakibara, Yasubumi, et al. "Stochastic context-free grammers for tRNA modeling." Nucleic acids research 22.23 (1994): 5112-5120.
[18] Tong, Simon, and Daphne Koller. "Support vector machine active learning with applications to text classification." The Journal of Machine Learning Research 2 (2002): 45-66.
[19] Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).
[20] 笹野遼平, and 黒橋禎夫. "大域的情報を用いた日本語固有表現認識." 情報処理学会論文誌 49.11 (2008): 3765-3776.