企業プレスリリースからのサイバーインシデント情報の自動収集と分析

池上和輝 †

明治大学総合数理学部　先端メディアサイエンス学科　菊池研究室 †

1　はじめに

近年，企業における内部不正や外部からの攻撃による個人情報漏洩などのサイバーインシデントが増加している．Japan Network Security Association(JNSA) は，インターネットニュースやプレスリリースからインシデント情報を収集し，発生企業・被害人数などの情報を人手で分類し，毎年報告を行っている [1]．しかし，我々が朝日新聞の記事データベース「聞蔵 Ⅱ」[2] を調査したところ，JNSA の報告にはない 2015 年のインシデントが 134 件報道されていた．情報収集に利用するメディアなどにより，報道されるインシデントに偏りがあることを示している．

そこで，本研究ではサイバーインシデントを，メディアなどによる偏りなく網羅的に自動で収集して分類することを目的とする．そのために，国内主要企業 537 社のウェブサイトをクローリングし，ウェブサイトのコンテンツに対して自然言語処理を用いることで，日付，被害人数，漏洩原因等のインシデント情報を自動で収集・分類するシステムを開発する．本システムにより，収集したインシデント情報の統計情報と分類精度を評価し，各種セキュリティマネージメント効果の検証に応用する．

2　クローラー・自動分類システム

2.1　クローラー開発

システムの全体構成を図 1 に示す．1. 企業ウェブサイトの URL 与える．2. ウェブサイトの html とそのページ内のリンクを収集する．3. 収集したコンテンツをテキストに変換する．4. 特定のキーワードを含むテキストを保存する．

キーワードには，[3] の調査結果に基づき情報漏洩に関するプレスリリースに含まれるであろう単語，「お詫び」,「漏えい」,「漏洩」を使用した．2 ではリンク先ページについても任意の回数繰り返し，URL を収集した各企業について 1-4 を繰り返し行う．

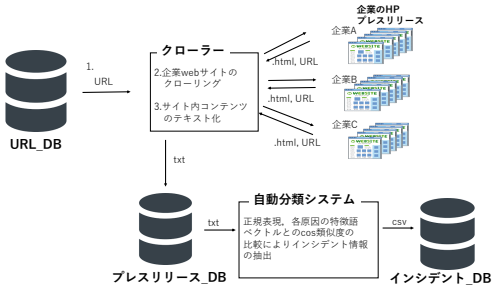

図 1　システム構成図

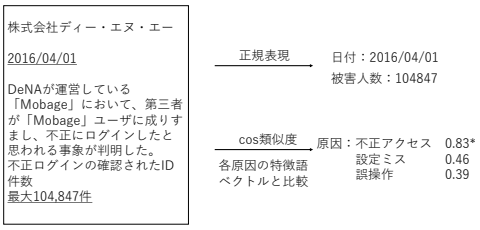

図 2　自動分類構造

自動分類システムの処理の例を図 2 に示す．日付，被害人数は正規表現により抽出し，日付は最も最新のもの，被害人数は最大の値を出力する．漏洩原因は，次の手順で推定する．1. 特定の単語の出現頻度とその単語を含む文書の出現頻度から定めた tf-idf 値を用いて，各漏洩原因の文章の特徴語を抽出する．2. 1 で抽出した各漏洩原因の特徴語から成る 49 次元の特徴語ベクトルを作成する．3. 未知原因のプレスリリースを参照し，2 で求めた特徴語の TF 値を求める．4. プレスリリースから抽出した特徴語ベクトルと各漏洩原因ごとの特徴語ベクトルの cos 類似度を求める．最も高かった漏洩原因を出力する．

2.2　調査対象企業

表 1 に調査対象企業の企業規模の内訳けと，セキュリティマネジメントの統計情報を示す．株式会社東洋経済新報社は，上場企業全社および主要未上場企業に調査票を送付し，その回答から社会的責任投資 CSR データベース [4] を作成している．本研究では各種セキュリ

†Kazuki Ikegami, Department of Frontier Media Science, School of Interdisciplinary Mathematical Sciences, Meiji University, Kikuchi Laboratory

表 1　調査対象企業の統計情報 (企業数)

企業数	大企業 2	大企業 1	中小企業	ISMS	CIO
537	193	225	119	169	150

表 2　クローラーの実験結果

期間	企業数	取得記事数	インシデント記事数	割合
2004/10/1 - 2018/11/2	537	17,957	191	(0.01)

表 3　推定結果

	日付	被害人数	漏洩原因	日付&規模&原因
適合率	0.882	0.792	0.719	0.505
	157/178	141/178	128/178	90/178

ティマネジメントの効果を検証するため，CSR データベース内の 537 企業を調査対象とした．537 企業のうち ISMS を導入しているのは 169 社 (31%)，CIO を設置しているのは 150 社 (28%) である．各企業の従業員数を元に，中小企業（従業員数 <300)，大企業 1（従業員数 <1500)，大企業 2（1500≤ 従業員数）の 3 種類に分類した．

2.3　実験結果

表 2 に実験結果を示す．2018 年 11 月，2.2 節の企業ウェブサイトに本システムを用いてインシデント収集・分類を行なった．取得記事のうちインシデントに関する記事は 191 件であり，全体の 1% であった．また，191 件にはインシデント詳細ページとインシデント一覧ページの重複があったため，一意なインシデントは 179 件である．

表 3 にインシデントの日付，被害人数，漏洩原因の推定精度以下の適合率で示す．

$$\text{適合率} = \frac{\text{正しく要素を抽出できたインシデント}}{\text{抽出した全インシデント}}$$

日付，人数などは全て 7 割を超えるが，全てを組み合わせた時の適合率は 50% であった．

3　取得したインシデント情報の評価

3.1　JNSA との比較

表 4 にインシデントを確認できた企業数とインシデント数を示す．比較対象として，JNSA データセットを使用する．公平に比較するために，表 4 では本調査結果を 2005 年から 2016 年の 34 企業，141 インシデントに限定する．企業数，インシデント数はどちらも JNSA の約 0.55 倍となった．また，クローラー独自で収集できたインシデントも，JNSA の件数の半分以下であった．

表 4　調査結果と JNSA との比較

	JNSA	JNSA 独自	本調査	本調査独自	共通
企業数	65	42	34	28	23
インシデント数	251	171	141	61	80

株式会社ディー・エヌ・エー

2016/04/01

DeNAが運営している「Mobage」において、第三者が「Mobage」ユーザに成りすまし、不正にログインしたと思われる事象が判明した。
不正ログインの確認されたID件数
最大104,847件

Jnsaとの共通インシデント

グリー株式会社

2016/12/27

グリーが運営するソーシャル・ネットワーキング・サービス（SNS）「GREE」およびスマートフォン向けゲームアプリを管理しているシステムに対して、不正なアクセスがあったことが判明

Crawler独自のインシデント

図 3　共通インシデントと独自インシデント例

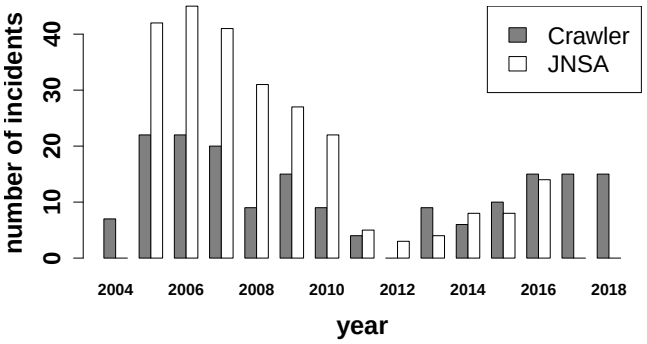

図 4　インシデント数の経年変化

JNSA 独自で収集できていて，本クローラーにより収集できなかったインシデント 171 件について，企業のウェブサイトを調査したところ，約 8 割にあたる 134 インシデントがウェブサイトに掲載されていないことを確認した．また 134 件のうち，インシデントの日付まで遡れないものが 40 件，日付は遡れるがウェブサイトに掲載がないものは 94 件であった．

インシデント数の年による変化を図 4 に示す．JNSA のインシデントは本調査の調査対象であった 537 企業に限定している．2005 年から 2012 年にかけてのインシデント数は，JNSA よりも少ないが，これはインシデントについてのリリースが HP に存在しないことが原因である．

図 5 に，インシデントの例と JNSA データセットと自動分類システムにより抽出した項目を示す．インシデントリリースを入力として，抽出できた取得項目を出力する．正解の出力は JNSA のデータセットにある項目とした．検出結果には，正しい検出と誤検出があるが，この

インシデントリリース(入力)

株式会社ディー・エヌ・エー
2016/04/01
DeNAが運営している「Mobage」において、第三者が「Mobage」ユーザに成りすまし、不正にログインしたと思われる事象が判明した。不正ログインの確認されたID件数最大104,847件

取得項目(出力)

	抽出結果	JNSA(正解)
企業名	ディー・エヌ・エー	ディー・エヌ・エー
業種	情報通信・サービスその他	情報通信業
日付	2016/04/01	2016/04/01
被害人数	104847	104847
漏洩原因	不正アクセス	不正アクセス
インシデント内容要約	未定義	有
参照記事のURL		
社会的責任度		普通
漏洩情報区分		個人情報
漏洩経路		インターネット
事後対応		普通

図 5　プレスリリースからの取得項目の違い

表 5　三つのデータセットのインシデント数の比較

	本調査	JNSA	聞蔵 II	共通
2015	10	8	3	2

例では全て正しく検出できている．インシデント内容要約以下の項目は未実装である．

3.2　聞蔵 II での調査結果との比較

表 5 に [3] と JNSA，本調査の比較を示す．[3] の調査では 2015 年分のみ収集したので，比較対象が一年分となっている．また，[3] による調査では 279 件のインシデントを収集したが，本調査対象の 537 社に絞ると 3 件のみになった．すべてに共通したインシデントは 2 件のみで，クローラーによるインシデント数が最も多い．

3.3　考察

インシデントの自動分類は適合率 50% 以下であった．これは，各要素の適合率で最も低かった漏洩原因の誤推定が原因だと考えられる．表 6 に判定の結果を示す．紛失・置忘れの誤推定のうち 75% が管理ミスと判定された．これは紛失・置忘れと管理ミスの tf-idf により抽出した特徴語が複数被っていたことが原因だと考えられる．

JNSA データセットとの比較の結果，本調査の 537 企業におけるインシデント数は JNSA の約半分であった．しかし，収集できていないインシデントの約 5.5 割は HP 上に存在せず，約 2.5 割は期間が古いため確認できなかった．2013 年，2015 年，2016 年のように近年のインシデントについては JNSA よりも多く収集できていた．

表 6　各漏洩原因の識別数

	紛失・置忘れ	管理ミス	盗難	誤操作	不正アクセス	ワーム等	その他
紛失・置忘れ	53	9	1	0	0	0	2
管理ミス	5	11	1	0	0	0	0
盗難	3	2	29	0	0	0	4
誤操作	0	0	0	12	2	0	1
不正アクセス	0	0	0	1	10	1	2
ワーム等	0	0	0	0	1	5	1
その他	0	1	0	0	4	2	8

4　セキュリティマネジメントの効果

本稿では，提案データセットの効果を確認するため，本調査によるデータセットと CSR データセットの突合を行い，ロジスティック回帰を行なった．本調査と [5] による結果を表 7 に示す．正の係数は（業種に当てはまる時，該当年の時，マネジメントを実施している）インシデントの生起確率が増加することを表す．例えば，本調査で正の係数となった情報セキュリティマネジメントシステム ISMS の認証を取った企業は，取っていない企業よりもインシデント数が多くなる傾向になる．同様に，最高情報責任者 CIO の設置，内部監査などのセキュリティ対策が [5] では負の係数になった．

5　おわりに

本研究では，537 企業のプレスリリースについてクローリングを行い．171 件のインシデントを収集し，JNSA，「聞蔵 II」によるデータセットとの比較を行った．収集したインシデントと JNSA データセットの比較結果，61 件の独自のインシデントが見つかった．インシデントの日付，被害人数，漏洩原因は 70% 以上の適合率で正しい分類ができたが，全要素を正しく分類する例は 50% 以下であった．

今後は，インシデンの識別精度を高め，対象の企業数を増やし汎用性の高い DB を提供することを課題とする．

参考文献

[1] 情報セキュリティインシデント調査報告書（JNSA データセット）

[2] 聞蔵 II，朝日新聞社，オンライン記事データベース．

[3] 池上和輝，菊池浩明，インシデント調査に基づく漏えい原因のデータマイニング, 情報処理学会第 80 回大会，2W-06， vol.3，pp.543-544，2018.

[4] 東洋経済データサービス CSR データ 2018.

表 7　本調査によるデータセットを用いたロジスティック回帰

		Estimate	Std.E*RR*	Pr(>|z|)	
a	(Intercept)	−23.260	2084.000	0.991	
b	建設. 資材	16.280	2084.000	0.994	
	素材・化学	16.950	2084.000	0.994	
	自動車・輸送機	−0.279	4188.000	1.000	
	鉄鋼・非鉄	−0.012	3548.000	1.000	
	電機・精密	16.260	2084.000	0.994	
	情報通信・サービスその他	18.120	2084.000	0.993	
	電気・ガス	20.330	2084.000	0.992	
	運輸・物流	0.367	14320.000	1.000	
	商社・卸売	0.467	3777.000	1.000	
	小売	17.540	2084.000	0.993	
	金融（除く銀行）	1.145	14510.000	1.000	
	機械	−0.219	0.065	1.000	
c	2014	−0.350	0.724	0.629	
	2015	−0.763	0.784	0.330	
	2016	0.752	0.595	0.206	
	2017	1.186	0.706	0.093	.
d	LOG(従業員数)	0.399	0.366	0.275	
x	ISMS	1.200	0.674	0.075	.
	CIO	0.000	0.719	1.000	
	内部監査	2.429	1.816	0.181	
	外部監査	−0.959	0.428	0.025	*
	内部窓口	−1.717	1.823	0.346	
	外部窓口	−0.969	0.789	0.220	

[5] 山田道洋，池上和輝，菊池浩明，乾考治，経営マネジメント状況による情報漏洩インシデント削減効果の評価 (2)，Computer Security Symposium 2018, pp.376-384, 2018.

表 8　[5] でのロジスティック回帰

		Estimate	Std.E*RR*	Pr(>|z|)	
a	(Intercept)	−8.300	1.072	0.000	***
b	建設・資材	0.223	0.800	0.780	
	素材・化学	−0.046	0.775	0.952	
	自動車. 輸送機	−0.334	0.981	0.734	
	鋼鉄・非鉄	−0.838	1.325	0.527	
	電機・精密	0.091	0.805	0.910	
	情報通信・サービスその他	0.561	0.738	0.448	
	電気・ガス	2.436	0.916	0.008	***
	運輸・物流	0.829	0.854	0.332	
	商社・卸売	0.066	0.849	0.938	
	小売	0.904	0.756	0.231	
	金融（除く銀行）	0.209	0.913	0.819	
	機械	−0.219	0.921	0.812	
c	2014	0.221	0.333	0.507	
	2015	0.185	0.343	0.590	
	2016	0.185	0.350	0.597	
	2017	−0.193	0.374	0.607	
d	LOG(従業員数)	0.948	0.255	0.000	***
x	ISMS	−0.217	0.331	0.513	
	CIO	−1.097	0.330	0.001	***
	内部監査	−0.207	0.374	0.580	
	外部監査	0.117	0.277	0.674	
	内部窓口	−0.050	0.761	0.947	
	外部窓口	−0.685	0.296	0.021	**