2W-06

情報漏洩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

池上和輝 †　　菊池浩明 †

† 明治大学総合数理学部

1 はじめに

業務の電子化に伴い，企業が保有する個人情報が漏えいするインシデントが多発している．例えば，2015 年 6 月に起きた日本年金機構の不正アクセスによる情報漏洩の様に，完全に防止するのは困難になってきている．そこで企業では，インシデントが生じたことを速やかに報じ，顧客の評判を落とさない先行対応にシフトしてきている．漏洩の損失額の評価には日本ネットワークセキュリティ協会 JNSA の JO モデル [1] が知られているが，迅速対応を評価するための事故から報道されるまでの日数情報が不足していた．

そこで，本研究では朝日新聞の記事から 2015 年の情報漏洩データセットを独自に作成し，漏洩原因についての連関規則から漏洩インシデントの特徴を明らかにする．漏洩件数と報道までの日数を分析し，企業の事後対応を評価する．

2 データセットの作成・分析

2.1 データセットの作成

本研究では，朝日新聞の記事検索システム「聞蔵」[2] を用いて 2015 年の漏えいインシデントの情報を収集した．「情報漏えい＋紛失＋漏洩＋不正アクセス＋誤送信＋盗難」の検索語を用いた．我々の調査では，JNSA では不足している，社内規則違反の有無や，報道時点での流出の可能性の有無，情報漏洩が起きてから報道されるまでの期間などの新しい要素を加えている．

表 1 に収集した漏洩インシデント数と JNSA を比較する．企業名と漏洩件数，公開日について両者が一致しているかを判断した．比較の結果の約半分が JNSA には含まれない新規のデータであった．逆に JNSA の件数が多いのは，JNSA のデータセットでは企業の支店まで分けてデータセットに加えていることが JNSA のデータセットの件数が多い要因の一つだと考えられる．

表 1　データセットのインシデント数の比較 [件]

JNSA	本データセット	共通
788	279	145

表 2　データセットの例

項目	JNSA	本調査	共通事例
公開日	2015/6/3	2015/9/7	2015/6/1
発生日*	不明	2015/9/2	2015/5/8
企業名	楽天証券	福岡県	日本年金機構
業種	金融、保険業	公務	公務
漏洩件数	1	35	1160000
漏洩原因	誤操作	紛失・置き忘れ	不正アクセス
漏洩経路	インターネット	紙	インターネット
漏洩要素	氏名	氏名/住所	氏名/ID
社内規則違反*	不明	0	1
被害の可能性*	不明	0	1
報道までの日数*	不明	5 日	24 日
社会的責任度など	有	不明	有

表 3　漏洩した個人情報レコード数の統計量

平均値	最大値	中央値	最頻値
8104	1160000	54	1

表 2 に本研究と JNSA データセットの例を示す．項目の*印は本研究のみの要素を示している．JNSA のデータッセットには，社会的責任度/インシデント内容要約/事後対応姿勢/経済的ランク/精神的ランク/基礎点/責任/対応/特定/一人当たりの損害賠償額が含まれる．

インシデントの公開日は新聞の発行日ではなく，記事中で述べられる公開日と漏洩日である．社内規則違反と二次被害については可能性が否定されていない事件を「有」と分類した．新聞の記事に「漏洩の可能性が極めて低い」，「誤って破棄した可能性が高い」，「一時的な紛失」などの記載がある場合は，二次被害の可能性が低いと判断する．

表 3 に収集した漏洩インシデントの漏洩レコード件数の統計を示す．最大値と平均値が高いが，中央値が 59.5 件で最小値が 1 件のことから，回数は少ないが規模の大きい事件が複数回あったことがわかる．最大値は 2015 年 6 月に起きた日本年金機構の不正アクセスによる漏洩事件である．それによりこの年の平均値が高くなったと考えられ，この事件を除いた平均の値は 3784 件と大幅に下がる．

Data mining of reasons of data breach based on the information leakage data set

†Kazuki Ikegami　†Hiroaki Kikuchi

† School of Interdisciplinary Mathematical Sciences, Meiji University

表 4　漏洩原因と漏洩要素の連関規則

No.	lhs	rhs	support	confidence	lift	件数	例
1	紛失・置忘れ	氏名	0.557	0.957	1.117	163	タカラトミー
2	誤操作	メールアドレス	0.079	0.595	4.757	37	愛媛県
3	管理ミス	氏名	0.089	0.926	1.080	27	長崎大学病院
4	管理ミス	住所	0.057	0.593	1.317	27	静岡ガス
5	不正アクセス	クレジット情報	0.014	0.190	13.334	21	日本年金機構
6	不正アクセス	ID/パスワード	0.014	0.190	7.619	21	新日本プロレスリング

2.2　漏洩原因と漏洩情報の連関規則

表 4 に漏洩原因と漏洩要素の連関規則および，各漏洩原因の件数，その代表的な事例を示す．「属性 A を持つ事例は属性 B を持つ傾向にある」という知識を連関規則という．lhs は条件，rhs は結論である．support は lhs と rhs の同時確率，confidence は条件付確率を表している．

また，lift は改善率といい，confidence が rhs の起こる確率の何倍かを示す [3]．この連関規則では，lhs に漏洩原因，rhs に漏洩要素を制約させることで，原因と要素から成るルールについて調べる．規則 1 は紛失・置き忘れがあると 0.95 の確率で氏名が漏洩することを表す．誤操作とメールアドレスの confidence が高いことから，それぞれの漏洩原因と要素に連関があることがわかる．また，規則 2 の lift4.75 とは，誤操作のときメールアドレスの漏洩が一般と比べて 4.75 倍生じやすいことを表す．lhs が誤操作のときのメールアドレスの漏洩という規則 2 の lift4.75 は，4.75 倍生じやすいことを表す．

2.3　報道までの日数

図 1 に報道までの日数のヒストグラムを示す．0 日から 9 日かかるものが最も多い．報道までに 10 日から 19 日かかる事件は 9 日以内と比べて約 4 分の 1 程度まで少なくなり，その後は緩やかに減少する．（また，報道までに 378 日かかった事件があったが，図をわかりやすくするために省略する）．

発生から報道までにかかる日数と漏洩した個人情レコード件数の関係を図 2 に示す．（報道までに 100 日以上かかった事件が 3 件あったが，図をわかりやすくするため省略する）．比較的規模の小さい漏洩ほど，報道までの時間がかかることがわかる．一方で，規模の大きい漏洩は会社の信用にも大きく関わるため，すぐに報道されると考えられる．また，平均日数は 11 日であったが報道までの日数の上位 3 件のみが 100 日を超えており，

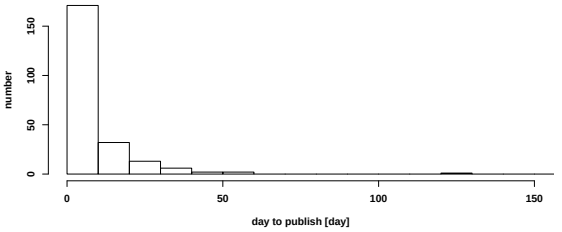

図 1　報道までの日数のヒストグラム

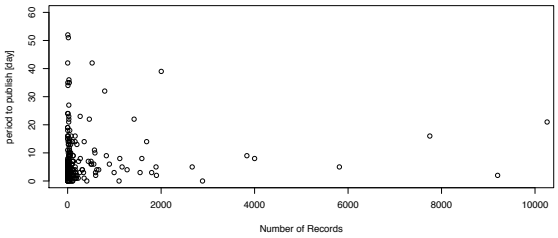

図 2　報道までの日数と漏洩の規模の散布図

それらにより平均日数が少し長めになったことが考えられる．

3　おわりに

本研究では，2015 年の新聞記事からデータッセットを作成し，漏洩原因についてデータマイニングを行った．連関規則により，いくつかの漏洩原因と漏洩要素に深い関係があることがわかった．また，規模の大きな事件ほど処理に時間がかかるため，報道までの日数は長くなると考えていたが，実際には規模の大きな事件ほど報道までの日数が短くなることがわかった．より大きなデータッセットを作成し、漏洩影響の評価をモデル化することを今後の課題とする．

参考文献

[1] NPO 日本ネットワークセキュリティ協会セキュリティ被害調査ワーキンググループ, 長崎県立大学情報システム学部情報セキュリティ学科, “2016 年 情報セキュリティインシデントに関する調査報告書個人情報漏洩編”, 2017.

[2]「朝日新聞記事データベース-聞蔵-」，(https://database.asahi.com/index.shtml, 参照 2017-12-21).

[3] 豊田秀樹，「データマイニング入門-R で学ぶ最新データ分析-」，東京図書，pp.147-183，2008.