3D1-R-13-1

企業内行動履歴解析による情報漏洩メール推定システムの検討

Examination of the data leak email estimate system by the action history analysis in the company

池田　利夫[*1]
The program committee of the XXth annual conference of JSAI

[*1] 関西電力株式会社
The Kansai Electric Power Co., Inc.

In this study, I predict the outbreak of the data leak email with an employee everyday behavior history in the company and suggest technique to prevent this beforehand. Employee actions to be able to put in companies such as WEB access number of times, business trip number of times, security training attendance time in the email transmission of a message number of times / a transmission of a message point number, PC use time. I sense the outbreak of the data leak email beforehand by making these information and causation with the data leak email clear by a multivariate analysis.

1. はじめに

近年，企業において，機密情報の社外漏洩が多発している．顧客情報などを含む機密情報の漏洩は，企業の社会的信頼を失うだけでなく，顧客に多大な迷惑を掛けることになる．

機密情報の漏洩防止は企業にとって喫緊の課題であるが，従来の防止方法では，その効果に限界がある．

例えば，最も情報漏洩経路として懸念される電子メールについて考えてみると，送信メールの件名や本文中に機密に関する用語があれば，システム側によって自動的に送信を保留するものがある．

これは，登録されている機密用語を含むメールについては，確実に漏洩を防止できるが，そうでないものについては，当然，漏洩することになる．また，チェックがシステム任せになるために利用者側のセキュリティ意識低下を招くことになる．

また，メール送信時に複数相手先アドレスが入力されている場合の警告メッセージのポップアップや，送信ログの管理も考えられ，これらは，一定の抑止効果を得ることができるものの，偶発的な漏洩メールの抑制に対してはあまり効果を期待することはできない．

その他，セキュリティ教育（啓蒙）の実施についても，教育受講後，継続的にセキュリティ意識を維持することは難しい．

関連研究を見てみると，論文[1]においては，ガボールフィルタと二次元同期符号の相関を用いて，印刷物の記載内容に依存せず，各種ひずみに強い電子透かしの検出を可能としている．

論文[2]においては，ユーザのプライバシーを保護しつつ不正コピーが実際に行われた場合にはその ID 情報が露見するデジタルコンテンツの配信プロトコルを提案している．

論文[3]においては，機密ファイル毎にラベルを付与し，ラベルに基づいて移動を制約することで情報漏洩を防止する手法を提案しており，また，論文[4]においては，企業情報システムにおいて，より強度な個人情報匿名化を実現するためのアルゴリズムを提案している．

これら関連研究では，機密情報が漏洩した際の追尾の容易性や，ラベル付けによる制限，より安全な匿名化などを目的にしたものであり，情報漏洩という観点では同じであるが，本研究の企業内行動履歴解析による情報漏洩メール推定方法とは異なるものである．

連絡先：池田利夫，ikeda.toshio@a3.kepco.co.jp

2. 手法

偶発的な情報漏洩メール発生は，以下のような企業内行動となんらかの因果関係があると推測する．

① メール送信
　メール送信件数，送信宛先数，送信時間帯，添付ファイル数，業務外用語記述など
② パソコン利用
　利用時間帯，利用時間数，印刷回数，ダウンロード回数，パスワード変更回数，セキュリティパッチ更新回数，利用場所など
③ Web アクセス
　閲覧禁止サイトアクセス数，Web メール送受信回数など
④ 勤務状態
　出張回数，会議回数，残業時間，社員属性（年齢，入社年），勤務形態（日勤，夜勤），旅費など
⑤ セキュリティ研修受講
　セキュリティ研修学習時間，セキュリティ研修テスト結果，社内セキュリティマニュアルアクセス件数など

実際に企業での顕在化した情報漏洩メール実績はほとんど存在しないため，解析に有意なサンプルを収集することは困難である．

したがって，今回，実際の顕在化した情報漏洩メールではなく，それに繋がる一歩手前の危険なメールを準情報漏洩メールと定義し，これと上記の企業内行動履歴との因果関係抽出手法を検討する．

準情報漏洩メールとは，ユーザからの明示的な苦情は発生していないが，個人情報（個人名，電話番号など）や機密情報（契約情報，金銭情報など）をメール本文や添付ファイルに記述して，大量に社外発信したものなどを言う．

これらのデータの解析から，例えば，以下，式(1)のような推測を行うことを目的とする．

（機密情報を含む多量社外メール発信回数）＝
　a×（出張回数）＋
　b×（残業時間数）＋
　c×（USB へのデータ出力回数）　　(1)

出張回数と残業時間数が増加するということは，業務が多忙になっているということを表している．

また，USB へのデータ出力回数が増えると言うことは，社外へのデータ提供が多く行われていることを表している．

これら「仕事が忙しい」「社外データ提供が頻繁」という状況に陥ると，メールを使った安易な機密情報を含む社外への大量送信を行う回数が増加する．すなわち準情報漏洩メール発生が増加することが予想される．

今回，実際にこれら企業内行動履歴データと準情報漏洩メールは様々な企業内制約上，取得することができなかったため，実データフォーマットに基づいたダミーデータを作成して，企業行動履歴データ抽出・編集処理，多変量解析処理，リスク判定処理，情報漏洩予測リスト出力処理など一連の処理が半自動的に実行されることを検証した（図 1）．

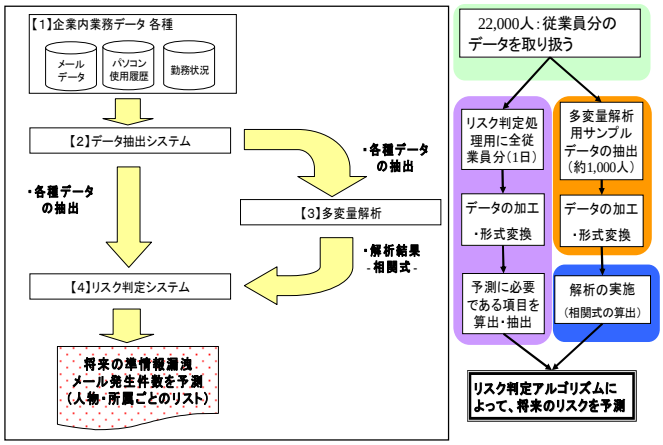

図 1：情報漏洩メール推定方法

3. 自動化の検証

企業内行動履歴および準情報漏洩メールのダミーデータを，約 22,000 人分を準備した．それぞれのデータ値は，実際に社内実データに即した値とした．

まず，これらダミーデータから多変量解析用データを抽出し，企業内行動履歴と準情報漏洩メールの因果関係式を，重回帰分析とロジスティック回帰分析により抽出した．

そして，この因果関係式をリスク判定処理に組込み，日々発生する企業内行動履歴の一定期間内データをこのリスク判定処理にかけることで，翌日，翌週，翌月に発生する準情報漏洩メールの発生件数を予測する一連の処理を実行することが可能であることが分かった．表 1 はリスク判定結果を表す．

表 1：リスク判定結果

	従業員No	判定日	翌日リスク判定結果（単位：通）			翌週リスク判定結果（単位：通）			翌月リスク判定結果（単位：通）		
			予測	前日実績	差	予測	前週実績	差	予測	前月実績	差
従業員別リスク判定結果	123456	2011/12/17	5	7	-2	13	10	3	28	30	-2
	111111	2011/12/17	4	5	-1	10	12	-2	22	19	3

	所属NO	判定日	翌日リスク判定結果（単位：通）			翌週リスク判定結果（単位：通）			翌月リスク判定結果（単位：通）		
			予測	前日実績	差	予測	前週実績	差	予測	前月実績	差
所属別リスク判定結果	654321	2011/12/17	10	9	1	18	17	1	36	32	4
	100001	2011/12/17	15	17	-2	20	21	-1	41	46	-5

日々，これら大量の社員データを用いて，予測処理（恒常処理）を実行する場合に，相当の処理時間を要することが想定される．データ加工や，メール添付ファイルのオープン，メールや添付ファイル内記述文書把握のための構文解析処理などが処理時間増加の要因である．

実際に測定した所，標準的スペックのサーバ処理で，約 7 時間かかった．長時間ではあるが，ユーザ利用頻度の比較的低い夜間バッチで処理を実行すれば，日々処理として，準情報漏洩メールの発生予測をユーザに提供できることが分かった．

また，多変量解析によって有意な因果関係式を抽出するために，目的変数（準情報漏洩メール）と複数の説明変数（企業内行動履歴）を総当りで自動解析する方法を行った．

重回帰分析において，説明変数が 15 個（総当り），サンプルデータ数が 1,000 個の場合，標準的パソコンで約 18 日間を要したが，これは，一連の処理が実行される前に一度だけ実行，または，1 年に一度程度実行すればよいものであるため（一時処理），長時間かかっても問題はない．

恒常処理（日処理）

処理	実測値
①データ抽出・整備	6時間
②リスク判定	1時間
①＋②	7時間

一時処理

処理	実測値
①データ抽出・整備，解析用ファイル作成	53時間
②多変量解析	432時間
①＋②	485時間（約20日）

図 2：処理時間

4. まとめと今後の予定

機械的で直接的に，情報漏洩メールを阻止する対策については，従来から様々な手法が考案されてきた．しかし，何れの対策においても，一定の効果はあるものの，偶発的・突発的に，無意識に発生する情報漏洩リスクに対してまで対策を実施することは困難であった．

今回の研究は，企業内での社員行動が，情報漏洩メール発生になんらかの影響を及ぼしているという仮定で研究を実施した．

大量に蓄積されている企業内情報の種類は様々であり，その膨大なデータの解析を実施することで，情報漏洩メールを発生させる社員行動パターンを特定できると考えられる．

ただし，その大量社内行動履歴データを自動抽出するには，大規模な基幹システムと連携した抽出システムを新たに構築する必要がある．

今回，実データを使用した解析を実施することができなかったが，今後は，社内基幹システムから一括して実データ抽出し，因果関係の解析を実施する予定である．

参考文献

[1] 前野蔵人他,:「情報漏えい対策に向く印刷文書用電子透かし方式」,電子情報通信学会発表論文,p.30-39,2007.

[2] 稲葉宏幸他,:「プライバシーと著作権を考慮したコンテンツ配信に関する提案」,電子情報通信学会発表論文,p.2536-2542,2006.

[3] 倉田兵武他,:「ファイル移動を制限する情報漏洩防止システム」,電子情報通信学会発表論文,p.186,2008.

[4] 佐藤嘉則他,:「識別リスクを保証する個人情報匿名化システムの検討」,DICOMO2007 シンポジウム発表論文,p.11821189,2007.