社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DD－64 (18)
2008－EIP－39 (18)
2008／2／1

個人情報の検出・分類技術とその活用・保護への応用方法

細見 格
NEC サービスプラットフォーム研究所

個人情報の漏洩を防止するための支援策として、個人情報を含む文書を自動的に検出するツールがある。しかし、従来のツールの殆どは人名や電話番号などを個別に検出するに留まり、本来の個人情報としての検出精度は低い。一方で、個人情報を正しく検出できれば、その漏洩防止だけでなく他にも有効な用途が考えられる。我々は、個人情報を高速に精度よく検出し分類できる技術を開発すると共に、その技術を情報セキュリティ管理におけるリスク評価に応用し、各文書中に含まれる個人情報の種類と数からリスクの評価に必要な資産価値を算定する方法を提案する。また、別の応用として、個人情報から問合せ先を区別することで Know-Who データベースを手軽に構築できることを示す。

Personal Information Detection Technique and its Effective Applications

Itaru Hosomi
Service Platforms Research Laboratories, NEC Corporation

We are researching and developing a personal information detection and classification technology. Such type of technology is usually used to protect personal information, but it potentially has other effective applications. We have considered such applications for the technology and found two applications, risk analysis for information security management and Know-Who database. This paper shows that these two applications can be realized easily by using our original personal information detection technique.

1. はじめに

2005 年 4 月より個人情報保護法が施行されて 2 年以上が経つ。同法施行の前から個人情報を保護するための様々な対策が論じられ、優れた効果を謳う技術や製品も数多く登場した。しかしながら、現在においても個人情報の漏洩事件は後を絶たない。

日米での情報セキュリティ調査の結果[1]からも、最近のノート PC などの盗難・紛失はセキュリティ事故全体の 3 割から 5 割近くを占め、情報漏洩防止のためにノート PC 内のデータを暗号化することが義務づけられている企業も多い。しかし、個人情報は人と人が互いを知り、コンタクトをとるために有用な情報であり、有用だからこそ蓄積される。従って、暗号化などの手段によって利用に不便が生じれば、ユーザはその制約を受けないところに個人情報を移して使うようになる。

暗号化のように強制的な保護手段をとる替わりに、個人情報の所在を明らかにしてユーザ本人に保護対策を促す方法もある。個人情報を含む文書ファイルを自動で検出するツールが各社から提供されているが、いずれも簡易な判定手法を採用しており十分な精度とは言い難い。加えて、ユーザにそのようなツールを使って貰うこと自体が大きな課題となっている。ユーザ自身に直接役立つ訳ではなく、場合によっては実行に数時間も待たされた後でその結果の確認が必要なソフトウェアの導入には、自ずと消極的になるからである。

我々は、従来製品と遜色ない速度でより高精度な個人情報の検出と分類が行なえる技術の研究開発を行なっている[2]。精度の面では従来技術に比べて大きく改善したが、それでも人が最終的な正否を確認し対策を決定する必要がある。検出精度を 100%にすることは事実上不可能なため、本技術を用いた情報漏洩対策の促進には別のアプローチが必要となる。

そこで我々は、個人情報検出・分類技術の応用について、100%近い精度でなくともユーザに負

担をかけずに使える用途と、ユーザに直接役立つ用途の2面から検討を行ない、前者として情報セキュリティにおけるリスク管理を、後者としてKnow-Who データベースを有望なアプリケーションの候補に選んだ。

以下では、まず我々の個人情報検出・分類技術についての概要を述べ、次にリスク評価のための資産価値算定への応用、Know-Who データベース構築への応用についてそれぞれ述べる。

2. 個人情報検出・分類技術

我々は、端末 PC 内に保存された文書中の個人情報を検出・分類するために、独自のレコード推定方式と機密文書オントロジを用いた手法を開発した[2]。本手法により、ファイル形式や文書構造に依存せず、高速に精度よく個人情報を検出することができる。更には、機密文書の種類や構成要素を定義した機密文書オントロジに基づき、個人情報を検出すると同時に自社員の連絡先や他社員の連絡先、問合せ先など複数の種類に分類することができる。

しかし、従来製品と同じく我々の以前の試作システムでは、検出結果の正否を確認するためには該当するファイルを開き、その内容をユーザが読む必要があった。そこで我々は、検出した個人情報の要素をその出現順序と要素間の位置関係に応じてシステム画面上に表示し、ユーザが別途ファイルを開くことなく検出された個人情報を見てその正否を容易に判断できるようにした(図 1)。

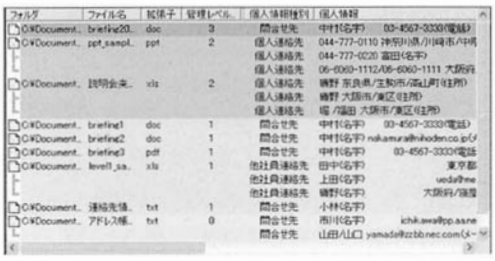

図 1 個人情報検出・分類結果の表示例

検出した個人情報は、次のような判定条件に基づいて各要素の配置を決定し表示している。

(1) 検出した1件ごとの個人情報の構成要素を、予め設定した幅に収まるよう、必要ならば各要素間の距離を同比率で縮小し1行に配置
(2) 距離を詰めても 1 件分の個人情報の構成要素全てを予め設定した幅の中に納まらない場合は、先頭から表示可能な分のみを配置
(3) 配置した要素間に十分な距離があれば、各要素の直後に、その要素の種類名(「住所」など)を付与
(4) 1ファイルから検出した個人情報のうち、重要度(後述)の高いものから指定件数分までを画面に表示(デフォルトは 3 件まで)

この表示方法では、要素単位での識別誤り(住所を人名と誤認識した場合など)の他、同じ行にある名前とメールアドレスの対応関係が正しくない場合(例: 山田 suzuki@foo.com)の誤りを判断し易くなっている。基本的に同情報を含んだ文書の所持者が見ることを前提としているため、覚えの無い組合せには気づき易いと考えられるからである。また、要素間の距離が遠いものほど実際には対応しない要素の組である可能性が高いとして、ユーザがより慎重に確認することもできる。

この改良により、検出結果の確認に掛かる時間を以前の表示方式(図 2)と比較したところ、1/3 程度に短縮でき、ユーザの利用負担を大きく低減できることを確認した(表 1)。

図 2 旧表示方式

表 1 表示方式の改良による効果の比較評価

	新表示方式	旧表示方式
自動検出・分類時間	19 分 14 秒	18 分 18 秒
人による確認時間	4 分 00 秒	12 分 50 秒

なお、本改良に伴う計算処理時間の差も比較したところ、表1に示した結果では約5%の速度低下が見られた。ただし内部的にはメモリ上の簡単な演算の追加であり、大幅な速度差は生じない。

従来の製品では、文書中に含まれる人名や電話番号それぞれの独立した数で個人情報の有無を判定しており、それら要素間の対応関係が考慮されていない。従って、個人情報を1件ごとに表示できず、上記と同様の効果を得ることが困難である。

3. リスク評価のための資産価値算定

以上のように、個人情報の検出・分類技術において、従来に比べてユーザの作業負担を大幅に軽減できた。しかし、検出結果の正否確認は、如何に容易であっても面倒であることに変わりはない。そこで、個人情報を検出・分類するツールを各ユーザの端末に導入して貰うが、各端末のユーザは確認作業等をしなくとも同ツールを有効活用できる方法について検討した。その結果、情報セキュリティ管理におけるリスク評価(図3)に適していると考え、その適用方法を具体化した。

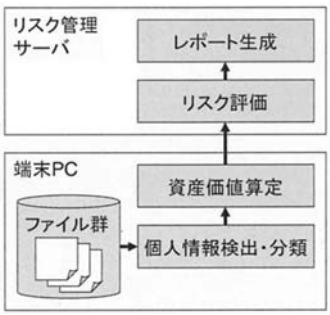

図3 リスク評価システムのイメージ

3.1. 資産価値の算定方法

ISO/IEC 17799 国際標準および JIS X 5080 国内標準となっている情報セキュリティ管理システム(ISMS)実施基準では、ISMS の運用に対する PDCA サイクル(Plan/計画, Do/実施, Check/監査, Action/改善を繰り返すサイクル)を規定している。この PDCA サイクルでは、最初の Plan の段階においてリスク評価(Risk Assessment)を必須としている。

セキュリティに関するリスク評価とは、一般的には以下の計算式に基づくリスク値を計算し、その値を予め与えられた基準値と比較することにより、問題点の有無とその深刻さを測ることである。

保有する資産のリスク値
＝資産価値×脅威の度合い×脆弱性の度合い
…(式1)

ただし、上式の演算子"×"は必ずしも乗算である必要はなく、資産価値、脅威の度合い、(資産保護の仕組みの)脆弱性の度合いそれぞれの算定方法についても、ISMS 実施基準や関連するガイドラインでは具体的に定義されていない。上記右辺の3つの要素はいずれも厳密な定量的評価が困難であるため、多くの場合それぞれ5段階程度のレベルで表現されている。即ち、最も価値の高い資産が最も脆弱な状態で保持され、最大レベルの脅威にさらされている場合、そのリスク値は5×5×5＝125(×を乗算とした場合)となる(0以上1以下の値に正規化される場合もある)。

我々は、企業資産の一種と言える個人情報の漏洩が、他の種類の情報資産の漏洩と比べて短期的には十分大きなリスクになると仮定した。そこで、情報漏洩に関するリスク値の算定において、個人情報の検出結果から資産価値のレベルを決定する方法を以下のような仮説の下で設計した。

(分類)情報資産＝{個人情報, その他の情報}
(仮説)
個人情報漏洩リスク >> その他の情報漏洩リスク
∴ 情報漏洩リスク値≒個人情報漏洩リスク値
∴ 情報資産価値レベル
≒個人情報の数と種類で決まる関数

現在、資産価値の算定に用いる個人情報の検出・分類結果は、その重要度(漏洩に対する危険度)の順に次の5種類を定義している。

(1) 個人連絡先(問合せ先を除く)
(2) 他社員連絡先(問合せ先を除く)
(3) 自社連絡先
(4) 問合せ先
(5) アドレス情報(一定数以上の電話番号またはEメールアドレス)

(4)の問合せ先は、その電話番号などにある程度不特定の人から連絡してもよいという意味で書かれた情報を指し、近傍に「お問合せ」などの文言を伴うもの、または文頭や文末の本文と離れた位置に1件独立して書かれた個人情報を対象としている。また、(5)のアドレス情報は、誰に対する連絡先かは不明でも大量のメールアドレスなどを含んだデータが流出すると個人情報漏洩の問題とされるケースが多いため設けている。

これら(1)～(5)の重要度別の個人情報を、さらにその件数によって2段階に分け、表2のように資産価値のレベルを割り当てた。

表2 個人情報の種類・件数別資産価値レベル

	1～S件	S件以上
個人連絡先	3	5
他社員連絡先	2	4
自社員連絡先 または問合せ先	1	3
アドレス情報	1	3

※閾値Sは、例えば100などの正数値

上記のような個人情報単位の資産価値レベルを用いて、1つの端末PCや1つの記憶装置といった一定範囲内に保持された情報全体の資産価値レベルWを次の式で求めることとした。

$$
\begin{aligned}
W = Max\{ & (個人連絡先件数 \geqq S)*5, \\
& (他社連絡先件数 \geqq S)*4, \\
& ((1 \leqq 個人連絡先件数 < S) \\
& \quad or\ (自社連絡先件数+問合せ先件数 \geqq S) \\
& \quad or\ (アドレス情報件数 \geqq S))*3, \\
& (1 \leqq 他社連絡先件数 < S)*2, \\
& ((1 \leqq 自社連絡先件数+問合せ先件数 < S) \\
& \quad or\ (1 \leqq アドレス情報件数 < S))*1\ \} \\
& \cdots (式2)
\end{aligned}
$$

ただし、"*"は乗算を表す演算子、Max{}は最大値を求める関数、各種類の個人情報や機密文書の件数に対する値域の制約は真ならば1、偽ならば0とする。

以上のような分類と式により、個人情報の検出・分類結果を用いた資産価値レベルの簡易な算定を自動で行なうことができる。なお、式1によれば、リスク値の計算には他に脅威と脆弱性のレベルをそれぞれ算定する必要がある。これらの自動化についても別途検討しているが[3]、ここでは詳細を省略する。また、従来のように既知の脅威や脆弱性を人手で分類してレベルを与えることは、膨大な文書から資産価値のレベルを算定することと比較すれば、まだ現実的と言える[4]。

3.2. 情報資産の重複の考慮

資産価値のレベルを文書に含まれる個人情報の種類と件数を基に算定する場合、その個人情報の重複を考慮する必要がある。1台の端末PCに含まれる文書や個人情報にはかなりの重複があると予想されるため、それらを単純に計上して前述の手法に基づく資産価値を割り出した場合、実際よりも大きな値になってしまうことが予想される。なぜならば、ある重要な情報を含むファイルを複製した2つの情報資産を同じ端末内に置いたところで、その端末単位での資産価値が増えるとは言い難いためである。すなわち、情報資産の価値は、互いに異なる情報資産の全体から割り出されるべきである。以下では、ある範囲内の情報全体に対する資産価値の算定において、式2による計算を行なう前に予め除去しておくべき情報資産の重複について考察し、その具体的な手法を述べる。

ここで対象としている情報資産の重複には、文書単位と個人情報単位との2種類が考えられる。ある文書の単純な複製、および編集した異なるバージョンの文書もその主たる内容に差が無ければ互いに重複した資産と見なしうる(文書単位の重複)。一方で、互いに内容の異なる独立した文書であっても、同じ重要な情報を含む場合もある(情報単位の重複)。異なる文書の執筆者が同一であり、その執筆者の連絡先や所属、役職などに関する個人情報が記載されている場合が、情報単位の重複の一例である。

なお、同一文書が何度も漏洩したケースと、内容が実質的に同じ別々の文書が漏洩したケースとは、資産価値やリスクの上では区別する必要が無いものと考える。なぜならば、いずれも同一資産が複数回盗難に遭うケースと同様であり、その資産価値自体は変わらず盗難事故の頻度の違いがリスク値に反映されるためである。従って、文書単位でも情報単位でも、それらの漏洩回数は資産価値に直接影響しないものと仮定する。

以上の考察から、まずは文書単位での重複を判定する方法を検討した。文書の類似性判定ではベクトル空間モデルを用いる方法が良く知られているが、そのためには各文書の単語ベクトル生成とその次元圧縮が必要になる。しかしそれらの計算量は、全文書の全内容同士の比較よりは少ないものの、各端末上でオンデマンドに行なうには大きすぎる。式1のようなリスク値の計算は専用の中央サーバ上で行なえば良いが、例えば個人情報を含むと判定されたファイル全てを評価対象の情報資産としてサーバに送ることは、サーバ側の負荷や端末側ユーザのプライバシーに対する懸念からも現実的ではないため、文書間の重複判定は各端末上で行なうか、計算負荷やプライバシーの問題の無い手法を採ることが望ましい。

そこで、個人情報を含むと判定された文書のファイル名同士の類似性と、検出された個人情報の種類ごとの件数を要素とした低次元のベクトル間での類似性を用いて、文書単位の重複を判定する手法を考案した。機密文書と判定された2つの文書ファイルの特徴ベクトル $\boldsymbol{di}, \boldsymbol{dj}$ の間の類似度は、コサインを用いた場合次のように与えることができる。

$$\cos(\mathbf{d}_i, \mathbf{d}_j) = \frac{[w_1 n_{pi}, w_2 n_{mi}, w_3 n_{oi}, w_4 n_{ci}, w_5 n_{ai}][w_1 n_{pj}, w_2 n_{mj}, w_3 n_{oj}, w_4 n_{cj}, w_5 n_{aj}]^T}{\|\mathbf{d}_i\|\|\mathbf{d}_j\|} \quad \cdots (式3)$$

w_1～w_5 は各要素の種類ごとの重み、n_{pi}, n_{pj} は個人連絡先件数、n_{mi}, n_{mj} は自社連絡先件数、n_{oi}, n_{oj} は他社連絡先件数、n_{ci}, n_{cj} は問合せ先件数、n_{ai}, n_{aj} はアドレス件数である。

上記のコサイン類似度とファイル名同士の編集距離を用いた類似度が共に所定の閾値を超えた場合に、比較した2つの文書ファイルは重複しているものと見なす。重複の除去は、ある2つの文書ファイルが重複と判定された際、双方の文書ファイルから検出された種類ごとの個人情報件数を比較し、件数の多い方のみをカウントすることで実現する。3つ以上の文書ファイルが重複した場合は、個人情報の種類毎に最大の件数分のみをカウントする。例えば、2つの文書ファイルAとBが重複と判定された場合、

文書Aから検出された個人情報:
個人連絡先5件、他社連絡先6件
文書Bから検出された個人情報:
個人連絡先4件、他社連絡先8件

であったとすると、結果として文書ファイルAとBに対する個人情報の合計件数は個人連絡先5件、他社連絡先8件となる。

あるホストマシンに機密文書として上記の文書ファイルAとBのみが検出された場合、式2より次のように資産価値レベルが決定される(閾値Sの値を例えば10とする)。

資産価値 W

$= Max\{(1 \leqq$ 個人連絡先件数$(\leqq 5)) < S) * 3,$

$(1 \leqq$ 他社員連絡先件数$(\leqq 8)) < S) * 2\}$

$= 3$

(※ 数量条件の判定結果が1となる種類の個人情報についてのみ記載)

各種類の個人情報件数に関する閾値が 10 であれば、重複を考慮しない場合は他社員連絡先件数の合計が 14 で閾値を超えるため、上記の資産価値レベル W の値は 4 となり、上記より1レベル高くなる。

さらに、文書単位での重複に加えて個人情報単位での重複も考慮することが望ましい。個人情報単位での重複は、第 2 節で述べたような検出結果表示に用いる個人情報の要素の組を互いに照合することで判定できる。ただし、互いに異なる種類の要素の組合せで一部のみ重複している場合に、どの程度共通していれば同一の個人情報と見なすかを決めておく必要がある。例えば、

(a) 鈴木, 03-1111-XXXX

(b) 鈴木, suzuki@foo.com, ○○株式会社

このような場合、(a)と(b)では人名しか共通項が無いため、たとえ同一文書から検出されたとしても重複とは判断しにくい。現在、我々は人名とその他1つの要素がいずれも共通している個人情報同士を重複と見なしている。

以上のようにして文書単位および個人情報単位での重複分を除去した資産価値レベルを計算することにより、少ない計算量で端末ごとの情報資産価値の近似値を得ることができる。資産価値レベルは 5 段階の離散値であるため、個人情報検出・分類の結果に含まれる多少の誤りは吸収することができる。現在の我々の個人情報検出・分類精度は 80%前後であり、資産価値レベルの判定であればユーザの確認なしでも実用になると見込んでいる。定量的な評価は今後の課題だが、さらに判定結果のサーバへの送信までを自動化すれば、各端末のユーザに負担無くリスク管理を行なうことができる。

4. Know-Who データベースの構築

徹底した処理の自動化によって、普及させたいソフトウェアに対するユーザの抵抗感を低減させる方法に対し、逆にユーザが積極的にそのソフトウェアを導入したくなるようなインセンティブを与えることも有効な手段である。

Know-Who とは、誰が何についてよく知っているのかを記録したデータベースであり、ナレッジマネジメント・システムの一部として提供されている場合が多い。ただし、従来の多くの Know-Who データベースは人手でデータを登録する必要があり、継続的なメンテナンスが課題となっていた。これについての対策の1つは、コミュニティ型のシステムであり、特定の人だけが登録するのではなく各ユーザが登録できる分散型の運用方法である。しかし、業務用の Know-Who データをユーザが忙しい仕事の合間に登録・更新していくことには、やはりある程度限界がある。

文書から個人情報を自動で検出する技術は、人による Know-Who データの登録作業を代替または補完する手段として利用できる。別途、各文書から主要なトピックやキーワードを抽出する手段が必要となるが、例えば予め製品名のリストがあれば、その製品名を含む文書から個人情報を検出することで、誰がその製品と関係しているかが分かる。

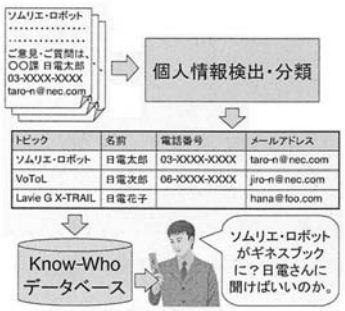

図 4 Know-Who データベースの構築・更新

また、先に述べた個人情報単位の重複判定と同様にして同じ人物の個人情報を同定し、それら同じ個人情報を含む文書と、同じ製品名を含む文書との積集合の数から、統計的にある製品と人物との相関性の高さを求めることも可能である。さらに、我々の個人情報検出・分類技術では問合せ先情報を区別することもできるため、あるトピックや製品について正規に問い合わせるべき人は誰かという判断も可能である。Know-Who データベースを幅広く共有する場合は、問合せ先情報のみを利用することが安全と考えられる。

いずれも精度の問題があるため、自動での検出・分類結果をそのまま Know-Who データベースに登録すると誤った問い合わせをしてしまう場合があるが、図 1 のような画面に抽出したトピック名や製品名も加えて表示すれば、確認しながら登録することができる。このようにして文書から Know-Who データベースを(半)自動的に構築するツールとしても使えれば、個人情報の検出・分類ツールをユーザがより積極的に導入できるものと期待している。

5. おわりに

Web(1.0)時代の到来により、デジタル情報の流通単位はファイルから URL で表されるページとなり、さらに Web 2.0 では RSS で配信されるブログなどの記事単位になってきている。ストレージ、フォルダ、ファイルなどは、いずれも情報の容れ物に過ぎず、次の時代には情報が意味の単位でアクセスされ利用されるものと考えている。

意味の単位で情報を識別するためには、情報に意味を表すメタデータを(タグ)を付与する方法や、意味単位の情報を抽出して情報源とは別にデータベース化(インデックス化)する方法がある。我々の個人情報検出・分類技術は後者の一例として位置づけることができ、個人情報を意味する要素の組を抽出することで、図 1 のように文書中にある各種の連絡先をピンポイントで確認することや、それを Know-Who データベースの一部として活用することもできる。

また、抽出した情報をまとめて統計処理を行なうことで、ある範囲(端末内のストレージ、特定のグループで共有されるデータ群など)の情報に対するセキュリティ上のリスクを評価することもできる。リスク評価への応用は、個人情報のように各端末上で抽出した意味単位の情報そのものは共有困難な場合にも、その統計量のみを集約して有効に活用する一例となっている。

あらゆるデジタル情報は、今後さらに共有され解析されていき、その有効活用と保護の両立がいっそう求められるようになると予想している。ブログや掲示板の書き込みを解析して評判を分析するような技術では、将来的には誰がどのような意見や傾向を持っているかまでデータベース化され、新たなプライバシー保護の仕組みが必要になるかも知れない。逆に情報漏洩を防止するための個人情報検出技術には、個人情報を安全に有効利用できる機能も同時に求められるかも知れない。そうした両立が可能な技術開発とシステム設計が、研究の段階からも必要になると考え、今後は本稿で述べた個人情報検出・分類技術の各応用方法について実効性を検証していく予定である。

参考文献

[1] 山口, 内田, 2007 年情報セキュリティ調査から見た情報セキュリティ状況の比較, 信学技報, ISEC2007-23, SITE2007-17, 2007.

[2] 細見, 情報資産価値と個人情報保護のための機密文書検出手法, 情処研報, 2006-DD-75, pp.53-60, 2006.

[3] H. Sakaki, K. Yanoo, R. Ogawa, A Model-Based Method for Security Configuration, Proc. of IWSEC2006, pp.60-75, 2006.

[4] 塚田, 企業を守るセキュリティポリシーとリスク評価, 日経 BP 社, 2001.