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

2006－DD－57 (8)
2006／9／29

情報資産管理と個人情報保護のための機密文書検出手法

細見 格
NEC インターネットシステム研究所

企業情報を安全に管理・活用するための情報資産管理で重要となる機密文書の自動検出について，サーバ側と端末側それぞれの要件に応じた手法を提案する．機密文書検出特有の技術課題に加え，多くの人が作成した大量の文書を少数の管理者が集中管理するサーバ側では人手作業を極力抑えるために高い検出精度が，屋外で持ち歩くするケースも多い端末側では特に個人情報の有無を持ち出し時に素早く確認できる高速性が，それぞれ重要な課題となっている．本論では，文書の構造解析と文解析の組み合わせ方，特に端末側個人情報検出用の表構造推定とオントロジによる情報分類の方式を中心に述べ，サーバ側と端末側双方のニーズに対する提案方式の有効性を示す．

Methods of Sensitive Document Detection for Information Asset Management and Personal Information Protection

Itaru Hosomi
Internet Systems Research Laboratories, NEC Corporation

We propose methods of server-side and client-side sensitive digital document detection. Information asset management has different requirements in its server-side and client-side point of views for the detection. The server-side requires full-automatic and high-accurate detection. While on the client-side, speed is most important factor especially for the handheld PC. We have met the both sides of requirements by two types of combination of document structure analysis and ontology-based text analysis.

1. はじめに

情報のデジタル化とネットワーク化の進展に伴い，組織内での情報の流通や共有が促進され価値ある情報の活用が容易になる一方，そのような情報の漏えいや改ざんの問題も急速に深刻化している．情報保護のための様々なセキュリティ対策も施されているが，その多くは保護すべき情報と公開可能な情報とを明確に区別しないため，資産としての情報の有効な活用をも制限する不便さがあった．そこで，利用を制限すべき情報の識別が，情報セキュリティと情報資産活用の両面から重要な課題となっている．しかしながら，一組織が有する情報資産は既に膨大であり，さらに日々変化，増大するため，人手による識別は不可能に近い．

特に最近は，個人情報保護に関する法律の施行や事件の多発から，個人情報の漏えい防止策が非常に重視されている．個人情報は，外部に流出すると大きな問題の発生に繋がりうるが，例えば企業において従業員や顧客を管理し，また各人が相互に連絡を取り合う上で必須な情報資産である．ここで，従業員情報などは主に組織内のサーバで管理されるが，連絡手段としてのアドレス情報は各自の端末で管理されている場合が多い．従って，個人情報の保護対策もサーバ側と端末側それぞれに対して適切に施す必要がある．

2. 情報資産管理と情報セキュリティ管理

情報のセキュリティに関しては，暗号化やアクセス制御などを用いた不正利用防止技術が数多く開発されているが，参照や流通を制限すべき情報とその他の広く公開/共有すべき情報との区別を明確にせずに，情報の利用制限だけを徹底させることは現実的でない．個人情報の保護と共に内部統制の重要性も叫ばれており，情報の安全性と透明性という相反するニーズをバランス良く満たす管理が求められている．これは，トータルな情報資産管理への要求と捉えることができ，情報資産管理の一側面として情報セキュリティ管理を考える必要がある．

そこで，本論では特に情報資産の保護に注目した情報資産管理の視点から議論する．情報資産管理の全体像は，前述の個人情報保護対策のように，大別してサーバ側と端末側から考えることができる．図 1 は，組織内の LAN 上に構築した情報資産管理システムの例を表している．図 1 のような構成において，組織内の情報は，その組織に所属する各個人が作成し各自の端末上に保管されているものと，各端末や組織外部から収集してサーバ上に保管されているものとがある．Eメールのように組織内外を流通する情報も，一時的には端末やサーバ上に蓄積されるため同様に扱う．

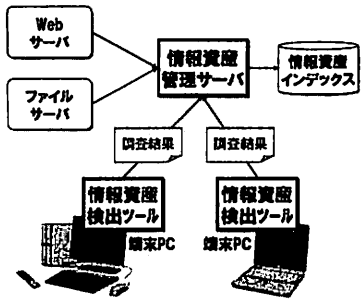

図 1 情報資産管理システムの概観

ここで想定している情報資産管理システムは，サーバ側では自身のローカルストレージや Web サーバ，ファイルサーバ等に蓄積された情報のうち有用な資産となる情報を判別し，インデックス化して管理する．端末側では，価値ある又は保護すべき情報を各端末上のツールや人手によって洗い出し，その結果をサーバに提供して上記インデックスに統合する．こうしてインデックス化された組織内の情報資産は，有用性や機密性の観点で分類することにより，その効果的な活用を促進するための情報ポータルを最適化したり，漏えいや紛失によるリスクを評価することができる．一方の端末側では，個人情報などを含む機密文書が見つかれば，その漏えいを防ぐように各端末で必要なだけの対策を施すことができる．

このような情報資産管理においては，蓄積されたデジタルデータから保護すべき資産となる情報を網羅的に検出する機能が求められるが，サーバ側と端末側ではそれぞれ表 1 に示したような異なる要件がある．

表 1 保護すべき情報資産の検出に対する要件

	サーバ側	端末側
実施形態	定期的	オンデマンド
実施頻度	低頻度	高頻度
重視される性能	高精度	高速
管理単位	ホストマシンやユーザ	ファイル
主な用途	リスク評価	情報漏えい防止

表 1 は，現状調査と識者へのヒアリングを基に各種の観点からそれぞれ典型的な要件を挙げたものだが，サーバ側と端末側との間で多くの違いがあることが分かる．サーバ側では，蓄積された情報資産の数や種類が膨大であり，サーバ管理者は個々の情報を評価できるだけの十分な知識を持たない場合が多い．従って，資産として保護すべき情報を人手をかけずに高粒度に識別できる性能が求められ，実際の管理は個々のファイルより粒度の荒いホストマシンやユーザの単位で行なわれると考えられる．

一方，端末側では，多くの場合個人の責任範囲にある情報が対象となり，ある程度は人手による評価も含めたファイル単位での管理が可能である．しかし，ノート PC のように端末自体を持ち歩くケースも多く，その際に端末ごと盗難に遭うことなどによる情報の漏えいや紛失の危険にさらされる．そこで，日頃の定期的な検査と共に持ち出し時のオンデマンドな検査が重要となり，その際にユーザを待たせないための高速性がツール側に求められる．

情報資産は様々な観点で分類されるが，以下では見つけ出して保護すべき情報資産の対象を，近年特に重要視されている個人情報を中心とした機密文書に限定して議論を進める．

3. 機密文書の自動検出

3.1. 対象とする機密文書の定義

「機密」とは，三省堂の大辞林第二版によれば，“[枢機に関する秘密の意]重要な秘密．主に政治上・軍事上の事柄についていう．”と説明されている．すなわち，「秘密」が公的・私的な非公開事項であるのに対し，「機密」は一般に公的(組織的)な非公開事項を指す．本研究においては，「機密情報」を“ある特定組織のメンバー間または特定の活動の関係者間でのみ共有すべき情報”とし，「機密文書」を“機密情報を含んだ文書”とする．ここで重要な点は，機密情報/機密文書は，第三者に無断で参照されてはならないだけでなく，当事者間で共有できなければならないということである．当事者でも参照困難な情報は，その資産価値が損なわれている．

機密文書の概念的な定義は上記の通りであるが，実際には非常に多様な文書が含まれる．秘密/機密には多くの場合時間属性があり，当初は非公開であってもある時点から公開情報となるケースは多い．また，どのような要素を含んでいれば機密情報と言えるかという一般的な定義も困難である．そこで，現時点では以下の2種類の文書のみを典型的な機密文書として自動検出の対象にしている．

(1) 機密ラベルが記載されている文書
(2) 個人情報を含む文書

(1)の機密ラベルとは，“取扱注意”や“秘密事項”など文書の先頭や末尾に記載される但し書きを指す．注意すべきは，本文中から“秘密事項”という文字列が検出されても，それは必ずしもその文書が秘密であることを示すとは限らず，秘密事項の取扱い方法に関する一般的な説明文かも知れない．また，丸秘マークなども機密ラベルに相当するが，画像として記載されているものは現在のところ対象外としている．

(2)の個人情報にも様々なタイプがあるが，ここでは最も典型的な個人情報として，人名や会員番号など個人を特定するキーとなる要素と，その所属組織名，住所，電話番号，Eメールアドレスといった連絡手段別

のアドレス情報が，互いに近接配置されているものを想定している．勿論，ある人物の名前とその連絡先が文頭と末尾のように離れて記載される場合もあるが，そのようなケースは比較的少ないと予想し，またそのような場合に相互の関連性の有無を判定するためには広域的な文書解析を要するため，次節で述べるように主に処理時間の制約から対象範囲を限定している．また，連絡先以外の身体的特徴などの個人情報も，網羅性な定義と評価の困難性から除外している．

3.2. 機密文書検出に対する制約

表 1 に挙げたような要件を含め，実際にサーバ側および端末側の環境で機密文書の自動検出を行なうには，以下に述べる種々の制約を考慮しなければならない．

(1) 検出単位

通常の情報漏えい対策では，ファイル単位で機密か否かを判定できれば十分だが，個人情報の数で文書の機密性をランク付けしたり，個人情報のみを隠蔽して文書を公開したい場合などでは，個々の個人情報単位で検出できる方が望ましい．

(2) 検出可能範囲

特に情報セキュリティの観点から，自動で検出可能な機密文書がどのようなものかを明確にすることが求められる．これは，組織内で定められた機密文書管理ポリシーとの整合性を確認したり，機密文書検出システムの導入効果がそのコストに見合うか否かを見積もる他，検出できない種類の機密文書に対して補完的な対策を検討する上で重要となる．

(3) 性能評価

機密文書検出システムの精度を評価する際，まず問題となるのが充実した評価用サンプルの入手である．検出対象が非公開文書であるため，実物を幅広く収集することは一般に容易では無い．また，検出対象を機密ラベルと個人情報に限定しても，それらの記述のバリエーションは表形式や自然言語文や特定の帳票に従うものなど，組織や個人によっても様々であるため，手作業で網羅的にサンプルを作成することも難しい．

(4) 対象ファイル形式

検査すべき文書のファイル形式は非常に多様であり，各社のワープロや表計算ソフトの専用形式，PDF，HTML，さらにそれらのバージョンによっても仕様が異なる．圧縮や暗号化を施されたファイルへの対応が求められる場合もある．

(5) 検出精度

サーバ側では人手による確認が困難なため高い精度が求められるが，ファイル単位のアクセス制御ではなくホストマシン単位のリスク評価などに適用する場合は，ファイル単位に 100%近い精度でなくとも実用になる．端末側でも，検出されるファイル数がユーザ自身で確認可能な範囲であれば，必ずしも高精度が絶対条件とはならない．ただし，逆に人手で対処可能な数に適切に絞り込む手段が求められる．

(6) 処理速度

前述したように，ノート PC のような端末上では数分程度で検査を完了できる高速性が求められる．一方のサーバ側では，専用ホストを設置できれば定期実施サイクル内に検査が完了すれば良い．他のサーバ・アプリケーションとホストマシンを共有する場合は，週末や深夜時間帯に完了する必要がある．

3.3. 従来のアプローチ

機密文書の検出は，技術分野としては情報（文書）分類もしくは情報抽出に相当する．分類技術として見た場合，機密文書と非機密文書に分類することになり，ルールやオントロジを用いる手法のほか，SVM (Support Vector Machine)[7]などの事前学習と統計的判定に基づく手法が考えられる．情報抽出技術として見た場合，個人の氏名や住所などの固有表現を抽出する問題は同分野において典型的なタスクであり，人名とその連絡先との組を1つの個人情報として抽出するタスクも概念検索や概念抽出などと称して良く知られた課題である[3]．情報抽出においても抽出対象の判定に SVM などが用いられるが，辞書やルールと組み合わせた手法が多い．

一方，個人情報を含む文書ファイルを自動検出するソフトウェアが既に幾つか実用化されているが，それらはいずれも人名や住所，電話番号などを辞書との照合で個別に検出している程度に留まり，本来の個人情報としての適合率は極めて低い．現状の製品の性能がそのようなレベルに留まっている理由は，前節で挙げた制約から次のようにまとめられると予想している．

(a) 機械学習を用いた手法の適用困難性
(b) 対象文書の種類の多さと網羅性の重視
(c) 処理速度に対する高い要求水準

(a) は前節の(1)(2)(3)に関係する．個人情報を1件単位で識別しながら機密文書を検出したい場合，さらには検出可能な文書の特徴を明確に定義したい場合，文書単位での学習と統計的判定基準に基づく分類手法は適用し難い．機密文書サンプルの入手困難性から機械学習では性能を出し難い場合も考えられる．(b) は主に前節の(4)(5)に関係し，特に情報漏えい対策としては，多様なファイル群から機密文書をなるべく漏らさず検出しなければならない．加えて，前節の(6)に基づく(c)の理由から，特に端末用では高速な処理のためにあまり複雑な処理方式の適用も困難となっている．

3.4. 本研究における基本アプローチ

本研究では，基本的には文書中から個人情報や機密ラベルを抽出する情報抽出の問題として捉え，また前節で述べた(a)～(c)の理由により，ルールやオントロジ（辞書）に基づく判定基準が明確でなるべくシンプル

な手法を採用することとした．ただし，従来方式とは異なり，キーワード抽出のみのテキスト解析だけでなく，キーワード間の関係や文書の構造的特徴を利用することで，自然言語文に限らず名簿などの表や列挙型の記述からでも必要な情報を精度良く検出することを目指している．

4. サーバ用機密文書検出システム

4.1. システムの概要

サーバ用に設計した機密文書検出システムのアーキテクチャを図 2 に示す．

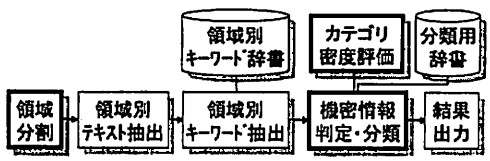

図 2 サーバ用機密文書検出システムの構成

サーバ用システムでは，まず読み込んだファイルを文書構造解析によって1つ以上の部分領域に分割する．次に，分割された個々の部分領域ごとにテキストデータを抽出し，各領域に対応する特徴定義辞書を参照して機密文書の判定材料となるキーワードを抽出する．さらに，各部分領域内でキーワード間の距離に基づく「カテゴリ密度評価」を行ない，機密情報の種類を判定する．以上の処理の後，機密情報の種類と機密度でソートして図 3 のような結果を出力する．

機密情報ファイル	主要機密文書	分類(管理レベル)	機密度
/usr/local/apache_1.3/htdocs/docs/	MANUAL5179.pdf	取扱注意	0.750
/home/hosomi/public_html/db/	bizcard.xls	個人情報(レベル1)	0.200
/home/hosomi/public_html/memo/	memo20040827.txt		0.000

図 3 機密文書検出・分類結果の表示例

4.2. 文書構造解析による領域分割

本システムでは，まず文書全体から構造的特徴を抽出し，その結果に基づいて文書を1つ以上の部分領域に分割する．この構造解析では，既存のソフトウェアモジュールを用いて Excel や PDF などの各種形式のファイルを HTML に変換(平文テキストはそのまま)した後，ヒューリスティックなルールを適用して文書からヘッダ／フッタ，タイトル，図表，図表のキャプション，表，および本文の各部分領域を抽出する．抽出ルールに合致しなかったテキストデータは全て本文領域と判断し，本文領域は規定文字数または規定行数(300文字または 5 行程度)ごとに分割してそれぞれを1つの部分領域とする．

領域分割の目的は，続く領域別キーワード抽出において，それぞれの部分領域の特徴に応じた辞書のみを参照することにより，辞書との照合処理を効率化し，また誤検知を低減する狙いがある．各領域別の辞書には，抽出すべきキーワードを機密情報のカテゴリ別にそれぞれ重み付きで登録している．例えば，ヘッダやフッタ領域で検出された“秘密”という語は，その他の部分領域で同じ語が検出された場合よりも当該文書が機密扱いであることを示す機密ラベルである可能性が高い．そこで，ヘッダ／フッタ領域用辞書に登録する“秘密”等の機密ラベル検出用キーワードには高い重みを与え，他の領域用辞書には同じキーワードを低い重みで登録するか，または登録しない．図表のキャプションなどでも部分的に機密扱いの情報であることを注記している可能性もあり，異なる部分領域で同じキーワードが検出されても，領域別辞書によりそれぞれ異なる重みを与えられるようにしている．

領域別辞書には，機密文書のカテゴリ毎にキーワードの組を記載している．例えば，「業務連絡先」カテゴリには人名，住所，電話番号，Eメールアドレス，さらに社名や部署名，役職などが含まれる．

4.3. カテゴリ密度評価による分類

領域分割と部分領域毎のキーワード抽出を終えると，各部分領域内において抽出されたキーワード集合に対し，それらが機密情報かどうかのより詳細な評価を行なう．

ある小領域内の人名や住所，所属名などは，それらが近接しているほど一連の個人情報である可能性が高い．構文解析による係り受け判定で要素間の相関性を評価することもできるが，名簿やEメールのシグネチャなどは自然言語文ではなく，構文解析自体の処理負荷も考えると大量の文書を対象とする場合は必ずしも得策ではない．そこで，特定のカテゴリに属する要素を含んだ最小の閉領域(カテゴリ領域)において，「カテゴリ密度」と称する値を計算することで，そのカテゴリに分類すべきかどうかを評価する方法を採用した [1]．

例えば，図 4 の上部破線枠内のような問合せ先の記述が文書中のある部分領域を構成していた場合，そこから抽出されるキーワードを列挙すると，図 4 の下部のようになる．

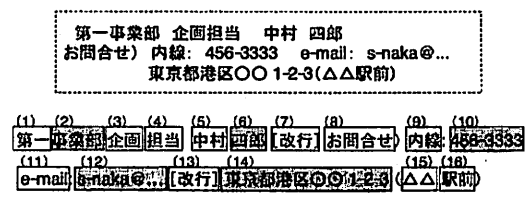

図 4「業務連絡先」カテゴリの要素例(グレー部分)

図 4 下部の各枠内が1単語(改行含む)，グレー部分が領域別辞書において「業務連絡先」というカテゴリに定義された要素とすると，カテゴリ領域は(2)〜(14)，カ

テゴリ密度はグレー部分の要素数/カテゴリ領域内の総要素数=8/13=0.615と計算する.

カテゴリ密度が規定の閾値未満の場合,そのカテゴリは分類候補から外す.閾値は全カテゴリに対して共通の定数としている.カテゴリ領域が重複する複数のカテゴリがある場合は,密度のより高いカテゴリに分類する.例えば,「業務連絡先」に比べて図 4 の(2)や(9)の要素を含まない「個人連絡先」というカテゴリがある場合,密度が「業務連絡先」より低くなるため分類候補から外れる.密度評価後も異なる複数のカテゴリが分類候補となった文書は,全ての候補カテゴリに属するものとする.

カテゴリ密度評価では,機密文書のカテゴリ分類を行なうと同時にそのカテゴリ毎の機密度を,検出した各キーワードの重みを基に算出する.ただし,機密文書間の比較は主に分類されたカテゴリと同カテゴリによって決まる管理レベルで行ない,機密度は参考値として出力している(図 3).

4.4. 実験と評価

サーバ用の試作システムは Xeon/3GHz の RedHat Linux 9 上に Perl をメインとして一部 C++ で実装し,サーバ上の約 16,000 ファイル(1.1GB)を対象に評価実験を行なった.機密文書として検出したファイル数は 192 であり,処理速度は 70.5KB/秒(5,130 ファイル/時),個人情報文書の適合率は 99.1%(正解 108 件,誤検知 1 件),機密ラベルで判断した個人情報文書以外の機密文書の適合率は 88.0%(正解 73 件,誤検知 10 件)という結果を得た.再現率は得られていない.個人情報の誤検知1件は,"名簿"というキャプションを含んでいたが実際には名簿ファイルへのハイパーリンクのみを持つ HTML 文書であった.また,機密ラベルの誤検知は,いずれも領域分割時にヘッダと判定した部分が実際には題名であり,その題名に"機密情報"などの語が含まれていた場合であった.

処理速度については,領域分割のために各ファイルから表やヘッダなどの構造情報を抽出する処理に,多くの時間を費している.また,一部 C++で実装することにより高速化を図っているものの,Eメールアドレスや住所の検出用に正規表現によるパタンマッチングを多用していることや,そのために大半を Perl で記述している点も速度に影響している.

4.5. 機密文書分類の応用

機密文書の検出・分類結果は,各文書のメタ情報として記録している(図 5)[2].このようなメタ情報は,前述のリスク評価に応用できる.

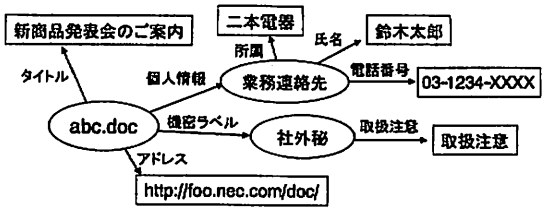

図 5 機密文書メタ情報の例

生成したメタ情報を図 1 における情報資産インデックスのようにデータベース化しておくことで,どのような種類の機密情報が組織内のどこにどの程度分布しているかを容易に把握することが可能になる.そこで,重要な機密文書が集中しているところからセキュリティ対策を優先的に施すような計画の策定を支援できる.

情報セキュリティ管理システム(ISMS)の標準的な手順では,組織内の情報資産を洗い出して分類し,その結果に基づいたリスク評価を行なうよう推奨されている[5].その実施には従来非常に多くの人的コストをかけていたか,もしくは洗い出す対象の情報量が多すぎて実施不可能であった.機密文書を自動的に検出し分類する技術により,リスク評価の前段階の作業コストが大幅に低減できる.

5. 端末用機密文書検出システム

5.1. システムの概要

先に開発したサーバ用の機密文書検出システムに対し,端末用では前述したように多くの面で要件が異なる.特に高速性への要求水準が高く,また OS も多くのバージョンに対応する必要があるため,サーバ用とは全く別にシステムを設計し,C++で実装した.

端末用機密文書検出システムのアーキテクチャを図 6 に示す.サーバ用に比べ,キーワード抽出を領域分割(レコード推定)より先に行なう点が構成上の主な違いである.ファイルからのテキスト抽出およびキーワード抽出には商用ソフトウェアを用いている.

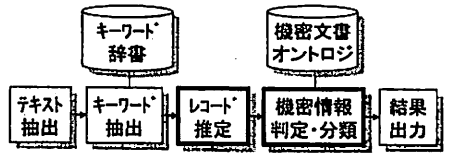

図 6 端末用機密文書検出システムの構成

本システムにおいて最も特徴的な部分は「レコード推定」機能である.その目的は,表形式で並ぶ個人情報の要素列を1件単位の個人情報レコードに分割することである.また,サーバ用のシステムではキーワード抽出に領域別のキーワード定義辞書,機密情報判定用に概念定義辞書,カテゴリ分類用にカテゴリ定義辞書(図 2 では概念定義辞書とカテゴリ定義辞書を分

類用辞書としてまとめている)といった複数種類の辞書を用いていたが，端末用のシステムではそれらを1つの「機密文書オントロジ」として階層構造で定義し，辞書のメンテナンスを容易にしている．

なお，端末用機密文書検出システムは，機密ラベルの検出も可能なものの，そのための機能はサーバ用のシステムに比べて非常に限定されており，実質的には個人情報検出専用となっている．

5.2. レコード推定方式

端末用機密文書検出システムでは，主に高速性とファイル形式への非依存性を確保するため，サーバ用のような文書構造解析を行なわない．文書構造解析では，各種の文書ファイルから構造情報を抽出するために，それらの様々なファイル形式の構造を個別に解析しなければならない．例えば，Microsoft の Office 2000 と 2003 との間でも幾つかのファイル構造上の違いがある．サーバ版では，個人情報や機密ラベルの検出精度を上げるためになるべく多くの構造情報を用いたが，端末版ではテキスト解析のみに限定することで，多少の精度低下と引き換えに対応可能なファイル形式の多さと高速性を得ることにした．

しかしながら，文書の構造的な特徴を使わずに文解析のみで個人情報を精度良く検出することは難しい[4]．これは，前述したように，多くの個人情報が名簿のような表やシグネチャのような語の羅列で構成される場合が多いためである．そこで，逆に全ての個人情報を表の要素列として捉え，表を構成する個々のレコードを推定して各々1件の個人情報として検出する「レコード推定方式」を提案する．本方式では，例えばシグネチャは唯一のレコードからなる表として検出される．レコード推定方式の概要を図 7 に示す．

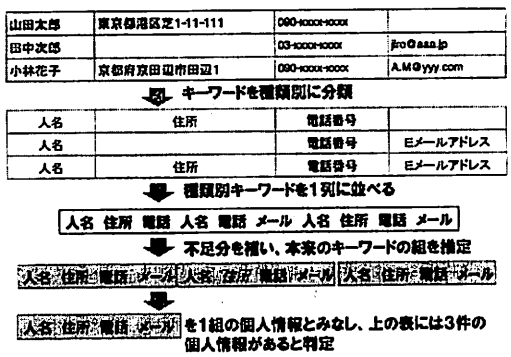

図 7 レコード推定の手順

個人情報は，実際に表形式で記述されていても，部分的に不完全な表である場合が考えられる．図 7 の最上部に記載した例では，1行目の最後と2行目の2項目めが欠如している．このような表の各要素から，個人情報の要素に該当するものをその種類の名前(例えば“山田太郎”ならば“人名”)で置き換えると，図 7 の上から2つめのような表になる．個人情報の要素に該当しないものは表から削除する．次に，表内の要素を全て1列に並べて先頭から参照していき，既に検出済みの項目が再び検出されたら，その直前までの一連の要素列を1つのレコードの候補とする．図 7 の例では，最初のレコード候補は{人名，住所，電話番号}となる．レコード候補に含まれる要素を除いた要素列から同様に次の候補を推定し，以前の候補との順序付き集合和を新たなレコード候補とする．2つめのレコード候補は{人名，電話番号，Eメールアドレス}となるが，1つめのレコード候補との集合和をとると，{人名，住所，電話番号，Eメールアドレス}になる．これを最後まで繰り返すと，表内の全ての個人情報要素がいずれかのレコードに含まれるようにレコードの要素列が推定される(図 7 の例では{人名，住所，電話番号，Eメールアドレス}となる)．なお，図 7 において，推定した各レコード内に斜字で書かれた要素(“*メール*”と“*住所*”)は，実際には対応する個人情報要素が無い部分を示している．

以上の手順で得たレコードの数が，その表に含まれうる個人情報の最大数となる．しかし，実際の個人情報の数は，各レコードに含まれる要素の集合を，次節で述べる機密文書オントロジと照合した上で決定される．

5.3. 機密文書オントロジ

本研究において，機密文書は個人情報もしくは機密ラベルを含む文書として識別しており，個人情報は，いわゆるプライバシー情報や個人宅の連絡先，所属する企業や学校の連絡先など，幾つかのタイプに分類できる．さらに，各タイプ(サーバ用機密文書検出システムにおいてはカテゴリに相当)の個人情報はその構成要素として人名や住所，所属する組織名などを含んで構成される．以上のような知識は，「機密文書」や「個人情報」といった概念を定義したオントロジとして表現することができる．これを機密文書オントロジと呼ぶこととし，端末用機密文書検出システムに実装した．

機密文書オントロジは，図 8 に示したような構造を持つ．基本的には対象の構成要素を記述したいわゆる part-of 関係の階層構造である．ただし，ある対象にとってその構成要素が必須か否かという観点から以下の3種類の属性を各2項関係に与えている．

MDT： 常に必須の構成要素
ALT： いずれか1つ以上は必要な構成要素
OPT： 必須ではない構成要素

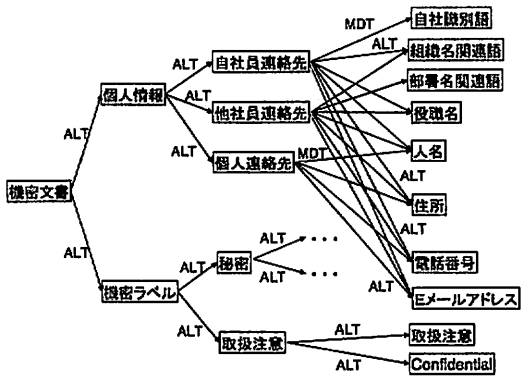

図 8 機密文書オントロジ

例えば，「個人連絡先」の構成要素として「人名」に MDT 属性を，「住所」と「電話番号」と「Eメールアドレス」に ALT 属性を，"教授"などの「役職」または「敬称」に OPT 属性をそれぞれ割り当てることができる．

ここで，前節の例で推定したレコードの構成要素は {人名，住所，電話番号，Eメールアドレス}だが，上記のオントロジに従えば，これらのうち必須要素の人名に加えて住所と電話番号とEメールアドレスのうちいずれか1つ以上があれば「個人連絡先」に相当する個人情報だと判断できる．すなわち，1レコード内に人名が無かったり他の3要素のいずれもが無い場合，そのレコードは個人情報としてカウントしない．機密文書オントロジでは，さらに人名を姓と名から構成されると定義したり，住所を都道府県名や市区町村名とその他の構成要素に分解した上で，都道府県名を実際の要素集合で定義することもできる．

機密文書オントロジは，文書から抽出すべき個人情報や機密ラベルをそれらの構成要素と MDT などの属性で明確に定義するほか，個人情報を「個人連絡先」「自社員連絡先」「他社員連絡先」といった複数のタイプに分類するための基準も与えることができる(図 9)．

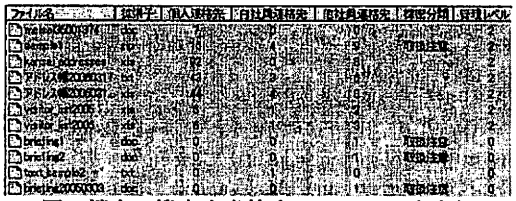

図 9 端末用機密文書検出システムの出力例

これら複数の用途に用いる知識を1つのオントロジとして記述できることで，同一要素を異なる辞書に別々に記述する場合に比べてそのメンテナンスが容易になるほか，組織内で定められた個人情報の定義や分類，機密ラベルの種類などを容易に反映できる．

レコード推定方式とオントロジによる個人情報検出は，表形式以外の場合にも有効に機能する．例えば，図 10 のような例からでも太字部分のキーワードの順列を抽出し，{人名，住所，電話番号}の組からなる5件の個人情報を検出できる．同様に，問合せ先なども1件の個人情報として検出できている．

山田さんは**東京都目黒区**に住んでいます．電話は **03-1234-XXXX** です．**加藤**さんは**世田谷区**在住で，携帯電話が **090-1111-XXXX** です．**中川**さんは**兵庫県明石市**にお住まいだそうですが，電話番号が分かりません．そのほか，分かっている方は以下の通りです．
藤井さん　奈良県生駒市　**0743-21-XXXX**
上田さん　名古屋市　**052-222-XXXX**

図 10 文章と列挙による個人情報記述の例

なお，1つの表とみなしてレコードを推定する範囲は，文書から抽出した隣り合うキーワードの間隔が規定値以内であるキーワード集合としている．キーワードの間隔が規定値を超えると，そこから先は別の表とみなす．同規定値はヒューリスティックに決定しているが，150 バイト程度で良い結果が得られている．

5.4. 実験と評価

端末用機密文書検出システムの処理速度は，Pentium4/3.2GHz の Windows XP 上で約 28,000 ファイル(7.8GB)を対象として実行した際に 1.1MB/秒となり，単純比較でサーバ用システムの 16 倍程度という十分な高速性を確認できた．また，従来方式によるキーワード抽出のみでの機密情報判定と比べても，速度低下率は 5%以下に留まっている．

精度についても，上記従来方式および同様の方式と推測される既存製品と比べて優位性を示すことができた(図 11)．

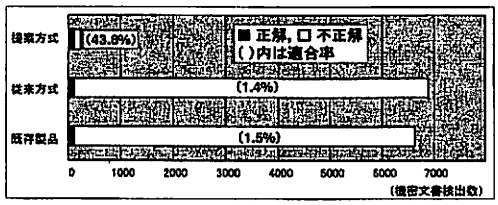

図 11　端末用機密情報検出の精度比較

人名やEメールアドレスなどいずれかのキーワードを検出するのみで個人情報と判定していた従来方式では，検出した個人情報の殆どが誤検出であったのに対し，その誤検出の 98%を除去できた．しかし，それでも適合率は 50%に満たない．これは，ファイル形式を限定せずに検査を実行した結果，テキスト抽出およびキーワード抽出において，EXE や DLL などのバイナリファイルから人名や地名に相当する文字列を頻繁に誤

検出していたためであることが判明した．評価対象のファイル形式を主要なオフィススイート文書やPDF，平文テキストなど13種類に限定した場合，適合率はファイル単位で100%，個人情報単位では95.5%となった．ただし，ファイル形式を限定した場合の処理速度は平均で667KB/秒となった．

なお，図11に示した実験結果では，予め用意した個人情報文書とその他の確認済み機密文書はいずれの方式でも全て検出結果に含まれていたが，ディスク上の全ファイルについては確認できていないため，サーバ用システムと同様に再現率は求められていない．

6. 考察

サーバ用機密文書検出システムでは，個人情報を含む文書で適合率99.1%という高い精度が得られ，一方で端末用のシステムでは適合率が下がるものの，1.1MB/秒という速度は実験に使用した文書1,000ファイルあたり5分未満で検査を完了できる．すなわち，初回はディスク上のファイル全てを検査するため時間がかかるが，2回目以降は検査済みファイルとの差分のみを検査するようにすれば，頻繁にノートPCを持ち歩くような場合，持ち出す直前の数分程度で新たな機密文書の有無を確認できる．

以上の結果は表1に挙げた要件をよく満たすものであり，サーバ用での領域分割とカテゴリ密度評価，端末用でのレコード推定といった各方式の有効性を示すことができた．特にレコード推定方式は，文書から抽出された一次元のキーワード列をストレートフォワードに評価する単純さ故の高速性が実証されたものと考えている．

ただし，サーバ用と端末用それぞれのシステムは殆ど独立に設計したため，今後互いに他方へ活用できる部分もある．サーバ用では，辞書に登録されていない人名も推定によって検出することで再現率の向上を図っている．例えば，表内のある列に人名が多く含まれていれば，人名が検出されなかった行の同列の文字列も一定の条件を満たせば人名と判断している．逆に，端末用システムで辞書をオントロジとして統一的に記述している点は，サーバ用システムにも適用できる．

技術的な課題としては，サーバ用では領域抽出のための文書構造解析やカテゴリ密度評価の方式がヒューリスティックな判定基準に依存しており，カテゴリ密度の閾値適性も十分評価できていない．端末側では，レコード推定方式に幾つかの問題が判明している．その1つは「レコードずれ」の問題で，表内の最初の個人情報要素が検出できないと，推定したレコードが全て1要素ずつずれてしまう．検出される個人情報数としては，レコードずれによる影響は表1つにつき1件少なくなる程度だが，個人情報を抽出して連絡用のアドレス情報に加えたい場合などでは，人名とその連絡先の対応関係が全て誤りとなるような大きな問題に繋がる．

また，これまでの実験ではサーバ用，端末用共に再現率を算出できておらず，意図した機密文書をどの程度網羅的に検出可能なのかを定量的に評価できていない．ただし，特に端末側の検出方式はファイル形式や文書の記述スタイルへの依存性が低く，検出されたファイルの種類からも機密文書の構造的バリエーションには幅広く対応できたと考えている．

7. まとめと今後

情報セキュリティ管理は情報資産管理の一側面として考えるべきであることを述べ，そこで重要となる機密文書の検出について要件を整理した．さらに，機密文書の自動検出システムについて，特に個人情報を対象としたサーバ用，端末用それぞれのアーキテクチャと検出方式を提案し，試作によって効果を確認した．現在，端末用の機密文書検出システムを社内で試用しており，ユーザからの意見を収集している．

今後は，レコード推定方式やオントロジの構築方法の改良によって，より多くの種類の個人情報に対応するほか，保護または活用すべき文書を特徴付ける個人情報や機密ラベル以外の抽出にも拡張していきたい．また，端末側とサーバ側双方から図5に示したようなメタ情報を収集し，情報資産のインデックスを構築することにより，リスク評価やその他の用途にも応用したいと考えている．

参考文献

[1] 細見 他，文書解析と設定検証に基づく情報漏洩脅威分析方式 (2)文書内容と構造解析を用いた機密情報分類，67th情処全大，3E-7，2005.

[2] 細見，情報セキュリティ分野におけるセマンティックWeb技術の活用，セマンティックWebコンファレンス2006.

[3] 松平 他，文書からのキーワード抽出と関連情報の収集，人工知能学会研究会資料，SIG-SWO-A303-02 (2004)

[4] 松本 他，表構造における意味的関係に基づくWWW検索精度の向上，情処研報，2006-DD-55，pp.5-11 (2006)

[5] (財)日本情報処理開発協会，ISMSガイド (Ver.1.0)，JIP-ISAC210-1.0，pp.6-12 (2004)

[6] http://chasen.naist.jp/hiki/ChaSen/

[7] C. Cortes and V. Vapnik, Support-vector networks, Machine Learning, Vol.20, No.3, pp.273-297 (1995)