3E-7

文書内容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

細見 格　　榊 啓　　小川 隆一
NEC インターネットシステム研究所

1. はじめに

最近のセキュリティ対策では，ポリシーと呼ぶ共通のルールを予め定義しておき，ポリシーに基づいて個別のセキュリティ設定や対処を自動化する試みがなされている．ただし，ポリシーを定義してセキュリティ対策を実施するには，保護対象を洗い出し，その種類や位置を把握する必要がある．特に近年は個人情報等の機密文書が保護対象として重視されているが，組織内に蓄積された膨大なファイル群から機密文書を人手で洗い出すことは不可能に近い．我々は，情報漏洩に繋がるセキュリティ設定上の問題を検出する技術の開発[1]において，独自の「設定検証ポリシー」を定義している．本稿では，このポリシー定義を容易にするために，機密情報を自動的に洗い出す機密文書検索･分類技術について述べる．

2. 機密文書検索・分類技術の必要性

「設定検証ポリシー」とは，どのような情報を誰が参照して良いのか／悪いのかをアクセス経路に関する条件を含めて定義したアクセスポリシーを，実際のファイル名やアクセス手段に具体化したルールである．例えば，「顧客情報を含む member.html は社外の者が Web サーバ S から http で参照できてはならない」というような内容である．設定検証ポリシーの定義には，どのような種類の機密文書がどこにあるのかを知る必要がある．デジタル文書は，ストレージ1つあたり数十万ファイル以上蓄積されることがあり，機密文書検索･分類作業の自動化は必須と言える．

3. 機密情報の分類手法

3.1. 情報検索技術による機密情報の検索と分類

従来のキーワード照合を基本とした情報検索技術によって，機密文書の検索や分類もある程度可能である．例えば，「取扱注意」といった語（機密ラベル）を含む文書や，住所･氏名･電話番号などを同時に含む文書を検索することで，それぞれ社内機密，個人情報といったカテゴリに文書を分類できる．しかし，「取扱注意」という単語の有無だけでは，本文中に「取扱注意とは～です」といった用語説明のみを含んだ文書も社内機密となる．また，互いに無関係な住所や氏名，電話番号が同じ文書中の離れた位置に検出された場合も個人情報に分類されうる．このように，特定種類の語の有無だけで機密文書の判定や分類を行なうと，多くの誤検出を伴う恐れがある．

3.2. 領域分割による機密情報検出の精度向上

「取扱注意」等の語がその文書全体を機密扱いとする機密ラベルなのか，住所や電話番号が特定個人の連絡先であるのかは，それらの語の書面上の位置関係や文脈に依存する．文書全体を機密とする場合，一般に文書や各頁の先頭または末尾に機密ラベルや但し書きを入れる．個人の連絡先やプライベートな情報は，個人を識別する名前や ID の近傍に記載されている場合が多い．このような空間構造上の特徴を利用することで，キーワード照合のみの方式よりも的確な機密情報の検出が期待できる．

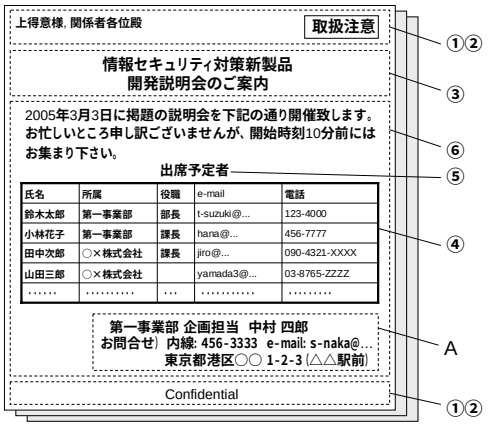

図1 文書の空間構造的特徴を示す例

そこでまず，文書を①文書全体のヘッダ／フッタ，②頁のヘッダ／フッタ，③タイトル，④表，⑤図や表のキャプション，⑥その他本文という7種類の領域に可能な限り分割する．「可能な限り」とは，文書の形式や内容によって必ずしもこれら全ての領域分割ができない場合を許容するためである．次に，それぞれの領域に対して適切な辞書を用いたパタン照合を行なう．パタン照合とは，ここでは機密文書のカテゴリ毎に定義されたキーワードの組と領域内テキストとの照合である．辞書には，カテゴリ名とキーワードの組の他に，各カテゴリの重要度（後述）を記述している．

An Information Leakage Risk Evaluation Method Based on Sensitive Document Detection and Security Configuration Validation (2) Sensitive Document Detection with Text and Structure Analysis
Itaru HOSOMI, Hiroshi SAKAKI, Ryuichi OGAWA
Internet Systems Research Laboratories, NEC Corporation

領域毎に異なる辞書を適用できるため，①②⑤のように複雑な文章が含まれない領域には小規模な辞書を使うなど，パタン照合の効率化や誤認識の低減を図ることができる．いずれかの領域がパタン照合によってあるカテゴリへの分類に必要な条件を満たした場合，そのカテゴリは対象文書の分類候補となる．

3.3. カテゴリ密度による要素相関性評価

氏名や住所，Eメールアドレス，年齢，カード番号などは，それらの関係が密であれば一連の個人情報である可能性が高い．自然言語文で書かれていれば構文解析による相関性評価も可能だが，名簿や名刺，Eメールのシグネチャなどは，ある範囲内に個人情報の要素が並んでいるに過ぎない．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提案する．

図 1 の本文領域⑥内の A で示した部分に含まれる要素を一列にして列挙すると，図 2 のようになる．

(1)第一 (2)事業部 (3)企画 (4)担当 (5)中村 (6)四郎 (7)[改行] (8)お問合せ (9)内線: (10)456-3333
(11)e-mail: (12)s-naka@... (13)[改行] (14)東京都港区○○ 1-2-3 (15)(△△ (16)駅前)

図 2 「組織情報」カテゴリの要素例（グレー部分）

カテゴリ領域は，領域①～⑥のうち 1 つの中で，あるカテゴリに含まれる要素が最初に見つかった位置から最後に見つかった位置までの範囲とする．図 2 の各枠内が 1 単語（改行含む），グレー部分が「組織情報」というカテゴリに含まれる要素とすると，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値未満の場合，そのカテゴリは分類候補から外す．規定値は全カテゴリに共通の定数としている．カテゴリ領域が重複するカテゴリがある場合は，密度がより高いカテゴリに分類する．例えば「組織情報」に比べ図 2 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「組織情報」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．

3.4. 文書の機密度判定

機密文書を分類すると同時に各文書の機密度を算出して比較することで，より機密度の高い文書に対する検査と対策を優先し，情報漏洩脅威分析の効率化を図ることができる．現在は，各領域について分類されたカテゴリのカテゴリ密度と辞書に記載された重要度との積を求め，その最大値を対象文書の機密度としている（図 3）．ここで，カテゴリ密度は機密情報である可能性の高さを，重要度はその情報が漏洩した際のリスクの大きさを表す値として用いている．

機密情報アドレス	主要機密文書	分類(管理レベル)	機密度
/home/hosomi/public_html/	visitor_list2004.xls	個人情報(レベル2)	0.800
/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 機密情報検索･分類結果の例

4. 機密文書分類システムの試作

以上述べた各手法を実装したシステムを試作した（図 4）．本システムは，指定した1つ以上のディレクトリや URL にある全てのファイルを参照し，機密文書の検出と分類，機密度に応じた順位付けを行なう．結果は図 3 のように Web ブラウザで出力されるほか，情報漏洩パス分析システム[2]のポリシー定義に利用するため，図 5 のような XML ファイルを出力する．

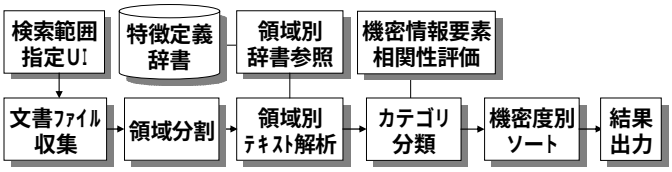

図 4 試作システムの構成

```
<result>
 <risk address="/home/hosomi/public_html/" value="0.800">
  <factor file="visitor_list2004.xls" score="0.800">
   <category>個人情報(レベル2)</category>
  </factor>
 </risk>
 <risk address="/usr/local/apache_1.3/htdocs/docs/" value="0.750">
  <factor file="MANUAL5179.pdf" score="0.750">
   <category>取扱注意</category>
  </factor>
 </risk>
</result>
```

図 5 機密文書とその分類結果の出力例

5. おわりに

大量の文書から機密文書を自動的に検索･分類し機密度を計算することで，重要な機密文書のある場所とその種類を特定する方法について述べた．これにより，実際に保有している機密文書に対する設定検証ポリシーの策定を容易にし，さらにその機密文書の機密度に応じて設定検証ポリシーを適用することで，設定検証[2]に基づく効率的な情報漏洩脅威分析が可能となる．

参考文献

[1]小川 他，文書解析と設定検証に基づく情報漏洩脅威分析方式 (1)コンセプトとシステムの概要
[2]榊 他，文書解析と設定検証に基づく情報漏洩脅威分析方式 (3)設定検証を用いた不正アクセス経路発見