情報セキュリティ管理における セマンティック Web 技術の活用

細見 格

NEC インターネットシステム研究所

1. はじめに

2005 年に入り、個人情報保護法が施行され、企業では情報セキュリティマネジメントシステム（ISMS）構築への取り組みが本格化するなど、安全な情報管理に向けた努力が続いている。しかし、未だ個人情報などの情報漏えい事故は後を絶たない。

ISMS では、情報を PDCA（Plan/計画，Do/実施，Check/監査，Action/改善）という4ステップのサイクルで管理することを推奨している。このうち、セキュリティ対策の実施（Do）については数多くのツールや機器が揃っており、その運用状況を監査（Check）するツールも市場に出始めている。しかし、情報セキュリティに関するリスク分析と対策方針策定を行なう計画（Plan）に有効なツールはまだ無いのが現状である。

上記のリスク分析では、まず会社などの組織が保有する情報資産を洗い出し、分類して種類別に価値を見積もる必要がある。これは、社内にある文書やデータを全て探し出して価値の高さでランク付けするという大変な作業である。その後、情報資産に対してリスク評価を行なうが、これも各情報がどんな種類の資産であり、どこにどのような状態で保管されているかも考慮すべきであるため、情報の柔軟な表現手段や評価手法が求められる。そこで我々は、情報資産のうち電子化された機密文書の洗い出しと分類を自動化し、その結果を柔軟に活用できるメタ情報として出力するシステムの研究開発を行なっている。

2. 機密文書探索・分類システム

我々は、ISMS の構築におけるリスク分析を支援するためのツールとして、機密文書の探索･分類システム（以下、本システム）を試作した[1]。本システムは、URL などのリストで指定した範囲の文書ファイルを参照し、自動的に個人情報または機密ラベルを含むファイルを検出する。ここで機密ラベルとは、「取扱注意」や「社外秘」など、配布制限のある文書の先頭やページの下端に記載される但し書きのことを指す。検出した機密文書は、さらに個人情報や機密ラベルの種類に応じて分類し、機密性のレベルで順序付けをした結果を表示する（図 1）。

ISMS では情報資産の価値を見積もることが要求されるが、本システムでは資産価値の代わりにその情報が漏えいした場合の被害の深刻度合いを指標とし、ヒューリスティックなレベルを割り当てている。

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 1 機密文書探索･分類結果の例

本システムの構成は図 2 のようになっている。機密ラベルが文面上のどこにあるかを判別するため、文書のレイアウト構造から領域分割を行なっている。また、リスク分析自体に関する機能は別のシステムとして動作し、本システムはリスク分析で用いる機密文書のメタ情報をデータベースに登録する。

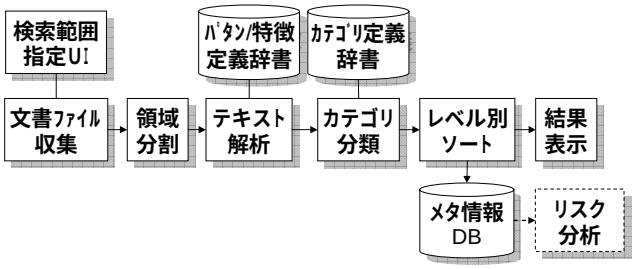

図 2 機密文書探索･分類システムの構成

3. 機密情報のオントロジによる定義

3.1. 機密情報の定義

機密文書を洗い出すためには、その文書に含まれる機密情報、ここでは個人情報と機密ラベルがどのような情報であるかを定義する必要がある。従来技術として、個人情報が含まれた文書と類似性の高い文書を探したり、単に人名や住所などが数多く含まれている文書を探すという手法はあるが、いずれも典型的な名簿ファイル以外にはあまり正確な結果が得られず、機密文書の分類にも適していない。そこで、本システムでは機密情報の要素構成を定義し、必要な要素同士および不要な要素との近接性を指標として機密情報の抽出と分類を行なっている[1]。

ISMS では情報資産の明確な定義が求められているが、同様に、情報資産（機密文書）を自動的に洗い出すシステムは、それがどのような情報を見つけられるのかを明確にすべきである。このことは、逆にシステムで自動的に見つけられないものを明確にし、同システムの導入効果や補完的な別の洗い出し手段を検討する際の判断材料を与える上で重要となる。

3.2. 情報抽出のための辞書

本システムでは、機密文書を検出するために次のような3種類の辞書を用いている。

1. 単語抽出用の文字列パタン定義辞書
2. 個人情報などを定義した特徴定義辞書
3. 機密文書を分類するためのカテゴリ定義辞書

実装上の辞書ファイルは更に数種類に分かれているが、概念的には上記の3種類に大別できる。

ある特定の特徴を持った文書を識別し分類するには、上記の辞書のような複数種類の知識が必要になることは明らかであろう。まず、文書中からある特徴の要素（辞書 1）となる単語や記号を見つけ、それらが見つけ出したい情報であるかどうかを特徴の定義（辞書 2）と照合して判断し、定義に合致した情報を含む文書を一定の基準（辞書 3）で分類する。これらの各ステップに応じた知識が必要となるため、自動処理を行なうシステムにも同様の各種知識を記述した辞書を備えなければならない。

3.3. オントロジによる定義の有効性

システムで用いる辞書が複数になることは、その辞書を記述するシステム利用者にとって負担となる可能性がある。多くの場合、上記の各辞書は相互に依存関係があり、具体的には一部に同じ用語を用いて記述される。そのような辞書が複数あると、定義すべき知識の全体像が把握しにくく、また記述ミスによって辞書間の用語の互換性が損なわれる可能性もある。

一般に、辞書というと定義する対象とその要素や説明の2階層で構成されるが、オントロジはより多くの階層構造や網構造によって、一般的な辞書よりも詳細で多様な知識を表現できる。上記のような3種類の辞書も、1つのオントロジとして記述することができる。

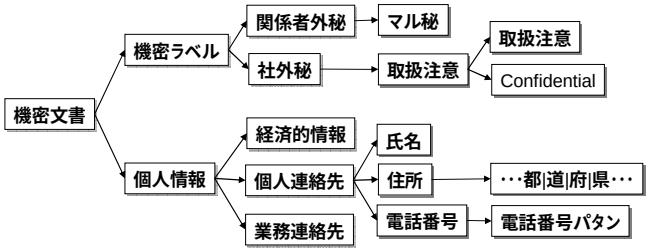

図 3 機密文書検出･分類のためのオントロジ記述例

図 3 に示したオントロジ（要素間の属性は省略）では、2～3階層めがカテゴリ定義辞書、3～4階層めが特徴定義辞書、4～5階層めが文字列パタン定義辞書にそれぞれ相当する。このように1つのオントロジで表現することで、利用する知識の全体が容易に把握でき、また要素の追加や変更もまとめて行なえる。抽出や評価の対象を明確に定義する必要があるシステムでは、オントロジによる定義が有効と考えられる。

4. 機密文書メタ情報の活用

本システムで検出し分類した機密文書に対して、抽出した個人情報などの要素や分類結果のカテゴリをメタ情報として書き出し、データベース（DB）に保存する。メタ情報は RDF グラフで表現している（図 4）。

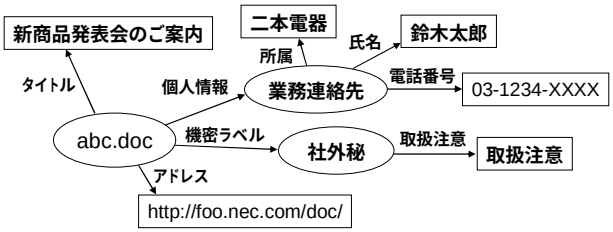

図 4 機密文書メタ情報の例

RDF の形式で DB 化することにより、一般的な表形式の DB スキーマに比べて新たな要素の追加や構造的変更が容易であり、同 DB を利用するシステム側の変更にも柔軟に対応できる。

生成したメタ情報は、別途開発中の情報漏洩脅威分析システム（以下、脅威分析システム）で利用される [2][3]。脅威分析システムは、社内のネットワークを通じて社外や他のドメインから機密文書が不正にアクセス可能かどうかを確認できる。上記のメタ情報によって、社内のどこにどのような機密文書があるのかが分かるため、機密性の高い文書が保存された領域が外部から容易にアクセスできる状態にあれば、リスクの高い問題箇所としてそのアクセス経路と共に指摘する。さらに、対策を施すべきポイントを優先度順に示すことで ISMS の計画の過程を効率化できる。

5. おわりに

ISMS の構築･運用者にとって負担の大きいリスク分析を支援するシステムを紹介し、その実現にオントロジとメタ情報を活用することの利点を述べた。セマンティック Web に関する議論では常にメタ情報やオントロジの作成が課題とされるが、本稿で述べたように、小規模なオントロジやスキーマ、自動生成可能なメタ情報を活かせる応用領域があるはずである。まずはそのような領域を探すことが大切であろう。

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]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, 67th 情処全大, 3E-6, 2005.

[3] 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, 67th 情処全大, 3E-8, 2005.