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

2008－DPS－134 (18)
2008－CSEC－ 40 (18)
2008／3／6

統計手法を用いた情報漏洩検知の改善に関する検討

竹口 誠士† 外川 政夫† 板倉 征男†

†情報セキュリティ大学院大学 〒221-0835 神奈川県横浜市神奈川区鶴屋町 2-14-1

あらまし

企業等においては、情報流出を未然に防止する様々な対策を実施しているが、通常、業務の中で機密情報（ファイル）がネットワーク上に流出しているかどうかの事実を把握する監視業務は実施していない。一方、インターネット上への機密情報の流出は、短期間に広範囲に情報が拡散するおそれが有るため、流出事実発見の遅れは被害を拡大する恐れがある。

そこで、著者等は、情報流出の検知プロセスを定義し、統計量によるファイルの特徴抽出とスコア表を用いて、同一及び類似ファイルの検知方法について提案を行い、その有効性を評価、検証してきた。

本論文では、その手法の中のファイルの特徴抽出の際のパラメータであるブロック長の最適サイズに関して検証したので、その結果を報告する。

キーワード 情報漏洩、統計手法、特徴量抽出、流出情報検知、スコア表

A study on improvement of the information leakage detection using statistics methods

Seishi TAKEGUCHI† Masao TOGAWA† Yukio ITAKURA†

INSTITUTE of INFORMATION SECURITY
2-14-1 Tsuruya, Kanagawa, Yokohama, Kanagawa, 2210835 JAPAN

Abstract

Almost enterprise companies implement the various measures which prevent information leakage, but they don't usually execute the network monitoring to grasp the fact whether the confidential business information (file) is leaked on the internet. On the other hand, leaked secret information (file) on Internet likely spreads to a wide area in a short period, therefore, the detection delay of leakage facts is a possibility damage expanding.

Then, the authors studied and proposed the information leakage detection process, statistical feature extraction from a file data, and a same (or similar) file detection methods using the score table. And we verified the effectiveness of the proposed method.

In this paper, as we verified in regard to the optimum size of the block length which is parameter in feature extraction of a file data, the result is reported.

Keyword information leakage, statistics methods, feature extraction, file detection, score table

1 はじめに

近年、インターネット等への情報流出事件が頻繁に起こり、情報の流出が社会問題になっている。

企業等においては様々な手法により外部への情報流出を防止する対策を講じているが、情報の流出事実について、把握が必要であるにも拘わらず、事実の把握に対する対策がほとんどされていないのが現状である。

そこで、著者らは流出の事実を把握するためのプロセスの提案及び類似情報の検知を可

能とする特徴抽出及び類似性の評価方法をSCIS2008で提案した。[1]

筆者等が提案した方法は、流出した情報と同一なファイルだけでなく、部分的に改ざんされたファイルについても発見することが可能である。そのために、ファイルの特徴抽出に統計量を用い、スコア表を用いて類似性を評価する手法を提案し、その手法の有効性について一定の成果を得た。

しかし、文献[1]ではファイルの特徴抽出の際に行うファイルの文字列データのブロック化について、ブロック長を1バイトに無条件設定して特徴量を算出しており、ブロック長を可変としたときの類似性に及ぼす効果については未検証であった。

本論文では、特徴抽出の際のパラメータの一つであるブロック長の最適サイズについて検証したので報告する。まず、第2章では、文献[1]で提案した統計手法を用いた外部流出情報の検知方法について概説する。第3章以下では、今回実験によって評価した特徴抽出の際の最適なブロック長について実験結果を踏まえて考察する。

2 提案手法の概要

2.1 情報流出の検知プロセス

文献[1]で提案している企業モデルは、大規模組織を想定しており、機密ファイルを保有・管理するファイル保有者とインターネットワーク上の流通ファイルを検索する探索担当者は異なる部門で実施することを想定している。大規模組織における情報流出の検知プロセスは、図1に示すように、3つのフェーズから構成される。

(1) 情報の特徴抽出フェーズ
ファイル保有者が特徴抽出を実施し、探索担当へ提出

(2) 探索・検知フェーズ
探索担当はインターネット上の流通情報を取得し、特徴DBをもとに流出ファイルの可否を判断

(3) 確認・分析フェーズ
探索担当は検出した同一あるいは類似のファイルをファイル保有者に渡し、事実の確認等を指示

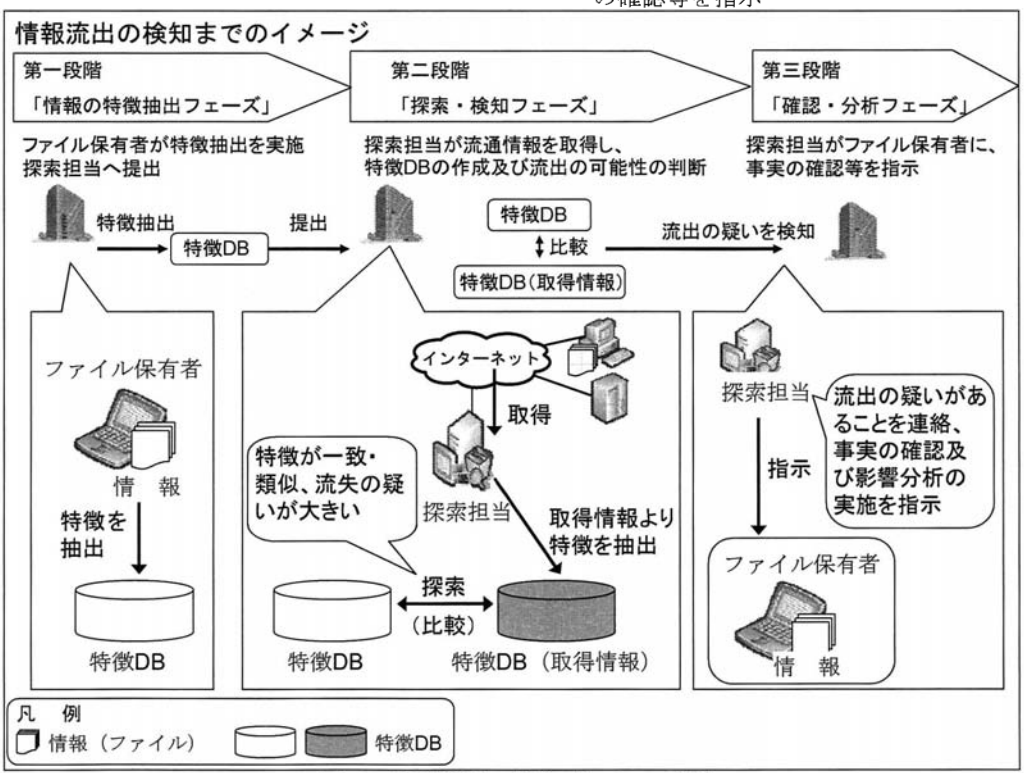

図1　情報流出の検知プロセス[1]

2.2 統計量を用いた特徴抽出

情報流出の検知プロセスにおいて、取り扱う機密ファイルは、次の 3 条件を満たすことを基本要件とする。

① ファイルには修正を加えない
(情報に電子署名やノイズ等を事前に埋込むようなものではない)

② 元ファイルを知ることが出来ない
(ファイルの取扱い権限が無い者への情報拡散防止が目的である)

③ 類似するファイルを認識できる
(情報の一部を変更・修正されたファイルについても発見したい)

統計量を用いた特徴抽出は、次の 2 段階のプロセスより行う。(図 2 参照)

【ステップ 1】
ファイルを 2 進数の文字列として取り扱い、その文字列を特定の長さ (ブロック長) で分割する。そして、その各ブロックを 10 進数変換した数値の母集合 (以下、標本と称す) を準備する。

【ステップ 2】
その標本で表されたファイルを統計計算により、「特徴量」で表現する。この抽出する「特徴量」は、「平均値」、「標準偏差」、「分散」、「尖度」及び「分布の歪度」の 5 つの「統計量」とする。

標本から直接算出される統計量は、統計学的には、観測 (観察) できるランダム変数の一種であり、『標本の性質を表現する数値』である。

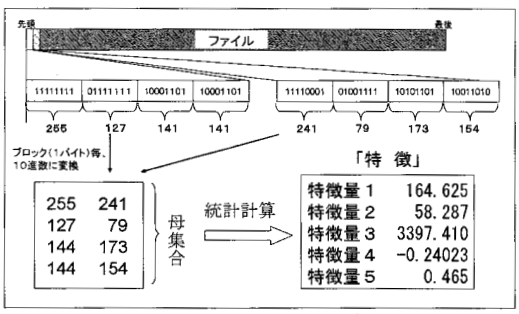

図 2 統計量を用いた特徴抽出 [1]

2.3 類似性評価の概念

著者等は文献 [1] において、類似性評価のために、「スコア表を用いた重み付け加算法」を導入することによって、類似するファイルの検知が可能であることを検証している。

以下に、導入した 2 つの指標を示す。

(1) 一致率 α

2 つのファイルの特徴量ごとの類似性を表わす評価指標として、一致率 α を定義する。すなわち、一致率 α は、5 つの特徴量がそれぞれ、どの程度一致しているかを表したものである。

検索対象ファイル (検索したいファイル) の特徴量を S、インターネット上から取得したファイル(被検索のファイル)の特徴量を X とすると、2 つのファイルの一致率 α は以下のように表わすことができる。

$\alpha_i(S)=(S_i/X_i)\times 100$ [i=1～5]

この時、$\alpha_i(S)$は検索対象ファイルの特徴量(統計量) i の一致率を示す。完全に一致していれば、一致率は 100 となる。

(2) スコア表

ファイルの文字数の変化量を変化率とすると、統計的な変化量で求めた一致率 α は、変化率と連動しないという特性がある。そのため、一致率 α だけではファイルの変化率を判断できない。そこで、あらかじめ既知の変化率を定めてそれに対応した一致率 α をスコア表の形で表して、それと、求めた一致率 α を照合させることにより、既知の変化率を基準に類似ファイルを推定できる。

本稿では、スコア表の作成に関して、以下の要件を定義する。

① 10%変化率までを類似ファイル と定義し、10%変化率を基準として類似ファイルを判定する

② ある基準ファイル (T) を定義し、基準ファイルが 10%変化した類似ファイルの一致率 $\alpha_{10\%}$ (T が 10%変化) をスコア 1 とし、α =100 の場合を 5 として、その間を 5 段階に均等分割してスコアを付与する。

ここで、10%変化した類似ファイルを様々な文字変換パターンで準備して、基準ファイル (T) との一致率 $\alpha_{10\%}$を求め、それらの平

均値を利用することにより、10%変化率の検出確率を向上させることが可能となる。

また、検索対象ファイルが一つの場合は、それを基準ファイル (T) としてスコア表を作成するが、検索対象ファイルが多数存在する場合には、検索対象ファイルごとに個別にスコア表を準備するのは困難なため、それらの中から代表的なファイルを複数用いてスコア表を作成することにより、多数の検索対象ファイルに対応がしている。

上記の2つの指標を使って、特徴量ごとに個別に求めたスコアを加算することにより類似性を評価している。

3 評価実験

3.1 評価項目

2.2節の統計量を用いた特徴抽出の【ステップ1】において、2進数の文字列で表現されるファイルをある一定のブロック長で分割し、各ブロックを 10 進数変換して得られる数値の集合を母集合と再定義している。

既に著者等が報告した文献 [1] では、このブロック長を文字の最小単位の1バイトに設定して検証した。

本論文では、このブロック長のサイズをパラメータとして変更した場合の類似性評価への影響及び最適なブロック長のサイズについて検証する。評価実験では、ブロック長のサイズを1バイトから4バイトまで可変して類似性の評価を試みる。

3.2 評価対象ファイル

評価実験に用いるファイルは、スコア表の作成に用いる基準文書と評価する検索対象ファイルからなる。実験環境を同じくするため、使用する基準ファイル, 検索対象ファイルは、文献[1]で用いた実文書をそのまま適用する。今回の評価実験に適用する実文書を表1、表2に示す。

表1 スコア表作成に用いる基準文書

法令名	略号
意匠登録令	意匠
官公庁施設の建設等に関する法律施行規則	官公庁
火炎びんの使用等の処罰に関する法律	火炎
官報及び法令全書に関する内閣府令	官報

表2 評価に用いる検索対象文書

法令名	略称	変化率%
関西国際空港株式会社法施行令 (全2499文字:189文字置換)	関西	7.563
国際観光事業の助成に関する法律 (全1393文字:176文字置換)	国際	12.635
水防法施工規則 (全2125文字:107文字置換)	水防	5.035

3.3 スコア表の作成

表1の基準文書の4法令を用いて、2.3節に述べた方法により以下の手順でスコア表を作成する。

【手順1】各基準文書の10%変化ファイルとして、次の5種類を用意する。その結果を表3に示す。

- 漢数字を算用数字に置換
- 漢字を異なる漢字に置換
- ひらがなをカタカナに置換
- ひらがなを漢字に置換
- 漢字を漢字、漢数字を算用数字に置換

【手順2】表3で準備した20種類の10%変化の基準文書も用いて、ブロック長を2〜4バイトの範囲で可変して、文献[1]の1バイトブロック長の場合と同様に、10進数変換した母集合を作成し、10%変化の基準ファイルの一致率 $\alpha_{10\%}$ を算出する。ブロック長2バイトのケースを表4に示す。紙面上、ブロック長3、4バイトは省略する。

【手順3】作成したブロック長2バイトの場合の一致率 $\alpha_{10\%}$ をスコア表にしたものを表5に示す。紙面上、ブロック長3, 4バイトは省略する。

3.4 類似文書のスコア計測

作成した表5のスコア表をもちいて、3種類の検索対象文書のそれぞれに対して、5%変化、7.5%変化、12.6%変化させた類似文書を作成して、それらの類似文書のスコア合計値を計測した。同様に、ブロック長3, 4バイトについてもスコア合計値を計測する。その結果を表6に示す。

ブロック長1バイトは、文献[1]で計測した値である。

4 考察

スコア表は、文書が10%以上変化している場合、スコアは原則的に0となるよう想定して作成されている。

ブロック長を1バイトとした場合は、文献[1]で検証した通り変化率 10%以下では変化率が小さい(則ち、類似性が高い)文書ほどスコアが高くなること、変化率10%を超えた文書はスコアが0となる(則ち、類似文書と言えない)ことを示しており、スコア表による重み付け加算法が類似文書の検知に有効であることを実証できた。

しかし、表6の実験結果をみると、ブロック長を 2～4 とした場合には、10%以上変化している3種の全ての文書において、本来0を示さなければならいのだが、スコアが4～8を示し、エラー判定となってしまう。また、10%以下の変化文書では逆にスコア0を示すエラー判定が生じている。

このように、ファイルの特徴抽出の際のパラメータであるブロック長の最適サイズは、文字の最小単位である1バイトとすることが最適であることが判明した。

これは、実文書を構成する文字列のコードの基本単位は1バイトであり、文字特性を保持した状態でコード変換（2進10進変換）するのが良いことを表している。すなわち、ブロック長を変更すると、コード変換した時点で文書の文字特性が変わってしまうことが要因として挙げられる。

5 まとめ

本論文では、統計手法を用いた外部流出情報の検知方法という新しい手法を提案し、その手法の中のパラメータのひとつであるブロック長について検証し、最適なブロック長を求めることができた。

今後、本検知手法を実際に応用していくためには、実験のサンプル数が少ないことや、検知精度の改善が求められることから、更に多くの様々な実文書を用いて検証を行っていく必要がある。

参考文献

[1]竹口誠士，外川政夫，板倉征男：統計手法を用いた外部流出情報の検知方法，SCIS2008，1B1-4，p44，（January 2008）

表3 実文書の総文字数等

法令名	総文字数	置換数	変化率	置換内容
意匠	2413	241	9.988	漢数字→算用数字241文字
意匠	2413	241	9.988	漢字→漢字241文字
意匠	2413	241	9.988	ひらがな→カタカナ241文字
意匠	2413	241	9.988	ひらがな→漢字241文字
意匠	2413	241	9.988	漢字→漢字118文字、数字123文字
官公庁	1401	140	9.993	漢数字→算用数字140文字
官公庁	1401	140	9.993	漢字→漢字140
官公庁	1401	140	9.993	ひらがな→カタカナ140文字
官公庁	1401	140	9.993	ひらがな→漢字140文字
官公庁	1401	140	9.993	漢字→漢字68文字、関数字→算用数字72文字
火炎	1039	103	9.913	漢数字→算用数字103文字
火炎	1039	103	9.913	漢字→漢字103文字
火炎	1039	103	9.913	ひらがな→カタカナ103文字
火炎	1039	103	9.913	ひらがな→漢字103文字
火炎	1039	103	9.913	漢字→漢字51文字、漢数字→算用数字52文字
官報	663	66	9.955	漢数字→算用数字66文字
官報	663	66	9.955	漢字→漢字66文字
官報	663	66	9.955	ひらがな→カタカナ66文字
官報	663	66	9.955	ひらがな→漢字66文字
官報	663	66	9.955	漢字→漢字33文字、漢数字→算用数字33文字

表 4　ブロック長 2 Byte でデータを取得した場合の一致率 $\alpha_{10\%}$

法令名 （置換概要）	2Byte				
	特徴量1	特徴量2	特徴量3	特徴量4	特徴量5
意匠 （漢数字→算用数字置換）	100.91248	100.55024	101.10351	101.43972	101.37754
意匠 （漢字→漢字置換）	99.95344	99.894755	99.789621	100.23389	100.17855
意匠 （ひらがな→カタカナ）	99.954868	100.05145	100.10292	99.615534	99.694998
意匠 （ひらがな→漢字）	99.185982	99.844026	99.688295	97.397774	97.74855
意匠 （漢字→漢字、漢数字→算用数字）	100.48816	100.25969	100.52006	100.90859	100.85139
官公庁 （漢数字→算用数字置換）	100.89329	100.60066	101.20494	101.16459	101.10715
官公庁 （漢字→漢字置換）	100.01255	100.02394	100.04788	99.955221	99.964065
官公庁 （ひらがな→カタカナ）	99.954835	100.03781	100.07563	99.669493	99.735926
官公庁 （ひらがな→漢字）	98.851275	98.9521	97.915181	99.57655	99.531053
官公庁 （漢字→漢字、漢数字→算用数字）	100.45648	100.29703	100.59495	100.63634	100.59079
火炎 （漢数字→算用数字置換）	100.60961	100.49135	100.98511	100.46006	100.50503
火炎 （漢字→漢字置換）	99.629725	99.367822	98.73964	101.0113	100.93217
火炎 （ひらがな→カタカナ）	99.745434	99.934845	99.869732	99.252806	99.328846
火炎 （ひらがな→漢字）	98.71813	98.998828	98.00768	98.884058	98.853592
火炎 （漢字→漢字、漢数字→算用数字）	99.950551	99.519843	99.041992	101.68145	101.5709
官報 （漢数字→算用数字置換）	100.92843	100.72571	101.4567	100.78465	100.83127
官報 （漢字→漢字置換）	99.904511	99.807907	99.616184	100.36933	100.35423
官報 （ひらがな→カタカナ）	99.955324	99.997175	99.994351	99.834841	99.853747
官報 （ひらがな→漢字）	98.815934	99.005491	98.020873	99.247488	99.204152
官報 （漢字→漢字、漢数字→算用数字）	100.48277	100.3055	100.61193	100.6805	100.71075

表 5　作成したスコア表（ブロック長 2 Byte ）

	特徴量1	特徴量2	特徴量3	特徴量4	特徴量5
スコア	範囲	範囲	範囲	範囲	範囲
0	$\alpha <$ 99.99	$\alpha <$ 99.94	$\alpha <$ 99.89	$\alpha <$ 99.79	$\alpha <$ 99.80
1	99.99 $\leqq \alpha <$ 99.99	99.94 $\leqq \alpha <$ 99.96	99.89 $\leqq \alpha <$ 99.92	99.79 $\leqq \alpha <$ 99.84	99.80 $\leqq \alpha <$ 99.85
2	99.99 $\leqq \alpha <$ 99.99	99.96 $\leqq \alpha <$ 99.97	99.92 $\leqq \alpha <$ 99.94	99.84 $\leqq \alpha <$ 99.90	99.85 $\leqq \alpha <$ 99.90
3	99.99 $\leqq \alpha <$ 100.00	99.97 $\leqq \alpha <$ 99.99	99.94 $\leqq \alpha <$ 99.97	99.90 $\leqq \alpha <$ 99.95	99.90 $\leqq \alpha <$ 99.95
4	100.00 $\leqq \alpha <$ 100.00	99.99 $\leqq \alpha <$ 100.00	99.97 $\leqq \alpha <$ 100.00	99.95 $\leqq \alpha <$ 100.00	99.95 $\leqq \alpha <$ 100.00
5	$\alpha =$100	$\alpha =$100	$\alpha =$100	$\alpha =$100	$\alpha =$100
4	100.00 $< \alpha \leqq$ 100.00	100.00 $< \alpha \leqq$ 100.01	100.00 $< \alpha \leqq$ 100.03	100.00 $< \alpha \leqq$ 100.05	100.00 $< \alpha \leqq$ 100.05
3	100.00 $< \alpha \leqq$ 100.01	100.01 $< \alpha \leqq$ 100.03	100.03 $< \alpha \leqq$ 100.06	100.05 $< \alpha \leqq$ 100.10	100.05 $< \alpha \leqq$ 100.10
2	100.01 $< \alpha \leqq$ 100.01	100.03 $< \alpha \leqq$ 100.04	100.06 $< \alpha \leqq$ 100.08	100.10 $< \alpha \leqq$ 100.16	100.10 $< \alpha \leqq$ 100.15
1	100.01 $< \alpha \leqq$ 100.01	100.04 $< \alpha \leqq$ 100.06	100.08 $< \alpha \leqq$ 100.11	100.16 $< \alpha \leqq$ 100.21	100.15 $< \alpha \leqq$ 100.20
0	100.01 $< \alpha$	100.06 $< \alpha$	100.11 $< \alpha$	100.21 $< \alpha$	100.20 $< \alpha$

表 6　類似文書とスコア計測値

略称	変化率 %	ブロック長：1Byte*		ブロック長：2Byte		ブロック長：3Byte		ブロック長：4Byte	
		スコア	判定	スコア	判定	スコア	判定	スコア	判定
関西	7.563	4	○	0	×	4	○	0	×
国際	12.635	0	○	6	×	4	×	8	×
水防	5.035	7	○	0	×	6	○	0	×

*文献[1]のデータ