D-019

正規表現・学習型フィルタ併用方式による機密情報検出の評価

Evaluation of a Confidential Information Detection Method with Regular Expression and Statistical Filtering

柴田 秀哉† 加藤 守† 郡 光則†
Hideya Shibata Mamoru Kato Mitsunori Kori

1．はじめに

近年の著しい情報量増加の中、電子データに機密情報が含まれるか否かを自動的に判別する機密情報検出の技術がその重要性を増してきている。しかし、既存方式では検出条件の作成が人手作業に大きく依存し、高精度な検出条件作成が困難であるという課題があった。そこで、容易に高精度な条件設定が可能な機密情報検出方式として、正規表現・学習型フィルタ併用方式を提案した[1]。

本研究では、正規表現・学習型フィルタ併用方式による機密情報検出精度の評価を行う。簡易な条件設定により機密情報がどの程度正しく検出されるかを、再現率、過剰検出率の 2 つの指標により評価した。結果、再現率 99.9%、過剰検出率 9.41%となり、正規表現・学習型フィルタ併用方式の有効性を確認した。

2．評価の想定

本評価では、評価データとしてメールデータを使用する。これは、機密情報検出技術の適用例の 1 つとして機密メールの検出システムを想定していることによる。これを踏まえ、本評価では社内メールを機密、社外メールを非機密として定義する。

3．評価方針

3.1 検証内容

検証項目 1 十分なデータ学習の下、学習型フィルタが機密情報検出フィルタとして有効に機能するか検証する。

検証項目 2 訓練用データの件数変化に伴う、検出精度の変化を検証する。特に、訓練用データ件数が少ないとき、正規表現フィルタとの併用効果が表れることを確認する。これは、システム導入時や検出条件の再生成時など、実際に機密情報検出技術が運用される場面において、データ学習が不十分な状況を想定している。

3.2 評価データ

評価データとして、業務メール 14,276 件を使用する。評価データにおける機密／非機密の内訳を表 1 に示す。

表 1：評価データの内訳

機密メール	12,575 件
非機密メール	1,701 件
計	14,276 件

業務メールを評価データとして使用するため、機密／非機密の件数に偏りがある。このような偏りは一般に起こり得ることであり、本評価においても偏りを持つデータセットをそのまま使用する。データの偏りに関する検証は検証項目 1 にて実施する。

3.3 条件設定

正規表現フィルタの条件は、検出対象文書の内容に精通していなくとも容易に設定可能であることを前提として設定する[1]。

本評価で使用したキーワード内訳を表 2 に示す。但し、検出条件は正規表現により記述され、単純なキーワードの羅列とはならないため、ここでは正規表現によって 1 語に括られる語群を 1 単語として数え上げている。

表 2：正規表現フィルタにおけるキーワード内訳

カテゴリ	設定単語件数
機密等級ラベル	1 件
定型文書名	171 件
定型文書登録番号	4 件
組織名略称	120 件
役職・人員名	182 件

3.4 評価指標

評価指標として、再現率と過剰検出率を採用する。併せて適合率による評価も実施する。再現率、過剰検出率、適合率の定義はそれぞれ次の通り。

$$\text{再現率} = \frac{\text{正しく機密と判定されたファイル数}}{\text{機密ファイル総数}} \quad (1)$$

$$\text{過剰検出率} = \frac{\text{誤って機密と判定されたファイル数}}{\text{非機密ファイル総数}} \quad (2)$$

$$\text{適合率} = \frac{\text{正しく機密と判定されたファイル数}}{\text{機密と判定されたファイル数}} \quad (3)$$

再現率は検出漏れの少なさ、過剰検出率は過剰検出の多さ、適合率は検出の正確さをそれぞれ表している。再現率と過剰検出率とは互いにトレードオフの関係にあり、機密情報検出では再現率が過剰検出率よりも優先される[1]。

3.5 交差検定

学習型フィルタのテスト方法として 10 分割交差検定を採用する。

4．結果と考察

4.1 検証項目 1

検証項目 1 の実験結果を表 3 にまとめる。この結果は、訓練用データとして被テストデータを除く全データ約

† 三菱電機株式会社 情報技術総合研究所
Information Technology R&D Center,
Mitsubishi Electric Corporation

13,000 件を使用したときのものである。従って、訓練用データの機密／非機密件数には偏りがある。

表 3：検出精度結果 (検証項目 1, 全データ学習)

フィルタ	再現率	過剰検出率	適合率
正規表現	88.0%	4.91%	99.3%
学習型	99.9%	5.21%	99.3%
併用方式	99.9%	9.41%	98.8%

表 3 より、学習型フィルタは十分なデータ学習の下、再現率 99.9%、過剰検出率 5%程度という高い精度を実現することが分かる。既存方式である正規表現フィルタの再現率が 88%程度に留まることからも、学習型フィルタの有効性が確認できる。

次に、訓練用データにおける機密／非機密件数の偏りが検出精度にどの程度影響を与えるか調べるため、検証項目 1 に関する別の実験結果を表 4 に示す。この結果は、訓練用データとして機密／非機密を 1,500 件ずつ、計 3,000 件を無作為に選定し、使用したときのものである。

表 4：検出精度結果 (検証項目 1, 訓練用データ選定)

フィルタ	再現率	過剰検出率	適合率
正規表現	88.0%	4.91%	99.3%
学習型	99.7%	2.31%	99.7%
併用方式	99.8%	6.86%	99.1%

表 3, 4 によると、学習型フィルタにおいて、訓練用データにおける機密／非機密件数の偏りにより検出漏れ件数が減少し、逆に過剰検出件数が増加している。これは学習型フィルタの判定結果が多数クラスに偏りやすい、という性質を表していると考えられる。学習型フィルタでは、訓練用データから特徴量を集計し検出条件を作成する。そして、未知文書の特徴量を検出条件と照らし合わせて判定する。従って、訓練用データ件数に偏りがあると、未知文書の特徴量が多数クラスに近いと判定されやすくなる。

本評価では、訓練用データ件数の偏りにより再現率が向上するという一見望ましい結果が得られる。しかし、常に機密が多数クラスとは限らず、非機密が多数クラスのときは再現率低下という結果となる。従って、一般には機密／非機密件数が偏らないような訓練用データ選定により検出精度は向上すると言える。

4.2 検証項目 2

検証項目 2 の実験結果を図 1, 2 にまとめる。

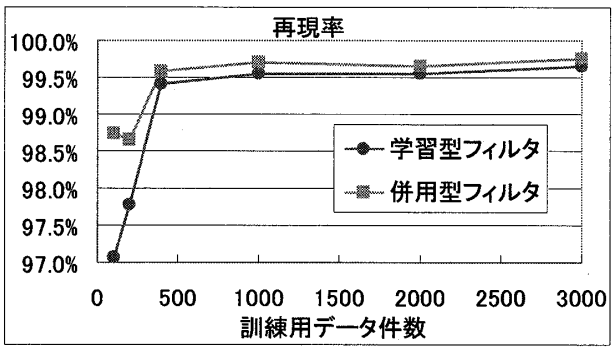

図 1：訓練用データ件数と再現率 (検証項目 2)

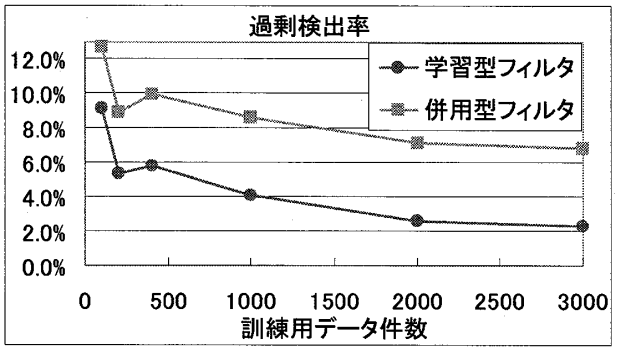

図 2：訓練用データ件数と過剰検出率 (検証項目 2)

この結果は、訓練用データとして機密／非機密を同数ずつ無作為に選定し使用したときのものである。図中の訓練用データ件数は、機密件数と非機密件数の和である。

図 1, 2 より、訓練用データ件数を増加させるほど、検出精度が向上することが分かる。また、図 1 より訓練用データ件数が 500 未満の少ないときに、併用方式が効果的に働いていることが確認できる。これにより、500 件程度の少ないデータ学習のみで十分な検出精度が得られることが分かった。

4.3 更なる精度改善に向けて

実験結果を分析した結果、学習型フィルタはサイズが小さいメールの誤判定を起こしやすいことが分かった。評価メールの中で 32 バイト以下のメールは全 7 件、うち 3 件が非機密であったが、表 3 の結果においてこれら 7 件は全て機密と判定された。サイズが小さいメールからは特徴量があまり生成されず、学習がうまく行えないことが原因と思われる。これは、サイズが小さいメールの誤判定だけでなく、その他の正しく判定されるべきメールの誤判定をも引き起こす可能性がある。

以上の問題への改善策として、サイズが小さいメールに対しては学習を行わず、判定の際は正規表現フィルタのみを適用するという方式が考えられる。この方式により、学習型フィルタによるサイズが小さいメールの誤判定が減少し、また、学習により生成される検出条件の精度低下防止につながる。結果として、更なる検出精度の改善が期待できる。これは、正規表現フィルタと学習型フィルタを用いるという提案方式独自の特徴を活かした方式である。

5．おわりに

本研究では、メールデータを使用した評価により、正規表現・学習型フィルタ併用方式による機密情報検出の有効性を確認した。簡易に設定可能な検出条件の下、機密情報を含んだ電子メールを高精度に検出することが可能である。十分なデータが学習されたときは、学習型フィルタにより高い検出精度が実現され、学習が十分でないときは、正規表現フィルタが検出精度を補完することにより安定した検出精度が実現される。

参考文献

[1] 加藤 他, 正規表現・学習型フィルタ併用方式による機密情報検出の提案, 第 8 回情報科学技術フォーラム, 5D-1, 2009.