MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

An Automatic Training Data Collection Method for Confidential E-Mail Detection

Hideya Shibata, Mamoru Kato, Mitsunori Kori, William Yerazunis

TR2010-065 June 2010

Abstract

In this paper, we propose an automatic method for operating a confidential e-mail detection system which uses machine learning and keyword search. The recent information explosion has increased the necessity of the technology which enables the detection of the confidential information in the electronic data. Using methods based on machine learning is one of the way for high accuracy. However, it is difficult to prepare a lot of correct training data manually, and this often becomes a problem for practice. We restrict our attention to e-mail, and present an automatic training data collecting method using the domain information. It allows the automatic operation of the confidential e-mail detection system. We also show the effectiveness of our method through the implementation and the evaluation for an e-mail archive system.

DEIM 2010

機密メール検出における訓練用データ自動収集手法

柴田 秀哉† 加藤 守† 郡 光則† William S. Yerazunis††

† 三菱電機株式会社 情報技術総合研究所 〒 247-8501 神奈川県鎌倉市大船 5-1-1
†† Mitsubishi Electric Research Laboratories 201 Broadway, Cambridge, MA 02139, USA
E-mail: †{Shibata.Hideya@cb, Kato.Mamoru@dn, Mitsunori.Kori@ab}.MitsubishiElectric.co.jp,
††yerazunis@merl.com

あらまし 本稿では, 機械学習およびキーワード検索を利用した機密メール検出システムの自動運用手法を提案する. 近年, 情報量の増大が著しい中, 人手による情報管理負担の低減を目的とし, 電子データにおける機密情報の自動検出技術が求められている. 機械学習に基づいた検出手法は高い精度が期待できる反面, 正確に分類された訓練用データを人手で大量に収集することは困難であり, このことがしばしば実用上の課題となる. 本稿では, 電子メールを対象とし, ドメイン情報を基にした訓練用データの自動収集手法を提案する. 提案手法により, 機密メール検出システムの自動運用が可能となる. また, 提案手法を電子メールアーカイブに実装し, 評価によりその有効性を実証する.

キーワード 訓練用データ自動収集, 機密メール検出, 電子メールアーカイブ

An Automatic Training Data Collection Method for Confidential E-mail Detection

Hideya SHIBATA†, Mamoru KATO†, Mitsunori KORI†, and William S. YERAZUNIS††

† Information Technology R&D Center, Mitsubishi Electric Corporation
5-1-1 Ofuna, Kamakura, Kanagawa, 247-8501 Japan
†† Mitsubishi Electric Research Laboratories 201 Broadway, Cambridge, MA 02139, USA
E-mail: †{Shibata.Hideya@cb, Kato.Mamoru@dn, Mitsunori.Kori@ab}.MitsubishiElectric.co.jp,
††yerazunis@merl.com

Abstract In this paper, we propose an automatic method for operating a confidential e-mail detection system which uses machine learning and keyword search. The recent information explosion has increased the necessity of the technology which enables the detection of the confidential information in the electronic data. Using methods based on machine learning is one of the way for high accuracy. However, it is difficult to prepare a lot of correct training data manually, and this often becomes a problem for practice. We restrict our attention to e-mail, and present an automatic training data collecting method using the domain information. It allows the automatic operation of the confidential e-mail detection system. We also show the effectiveness of our method through the implementation and the evaluation for an e-mail archive system.

Key words Automatic Training Data Collection, Confidential E-mail Detection, E-mail Archive

1. はじめに

近年, 情報量の増大が著しく, 文書に含まれる機密性の判断を個人に委ねると故意や誤りによる情報漏洩を防止できないという問題がある. 日本ネットワークセキュリティ協会が出した調査報告によると, 2008 年に起きた情報漏洩事故は 1,373 件に上り, 前年の 2007 年と比較して 509 件増と増加傾向にある [1]. このような背景があり, 人手による情報管理負担の低減を目的として, 電子データに含まれる機密情報を高精度に自動検出するための技術が求められている.

機械学習に基づいたテキスト分類手法は, 人手による条件設定なしに高精度な分類を可能とするため, 機密情報検出においても機械学習を利用することで高精度な検出が実現されると期待できる. 実際, 筆者らは容易な条件設定の下に高精度な検出が可能な機密情報検出の方式として, キーワード検索と機械学習とを併用した正規表現・学習型フィルタ併用方式を開発し, そ

の有効性を評価実験により示している [2], [3].

しかしながら, 一般に機械学習を使用する場合, 検出精度は訓練用データの与え方に大きく依存するため, 正確に分類された訓練用データを大量に収集する必要がある. 十分な検出精度を確保するために必要とされる訓練用データ件数は数千件以上に及ぶこともあるが, 多くの場合, このデータ収集作業は人手で負担することとなり, このことがしばしば実用上の課題となる. 特に, 機密情報検出技術を人手を介さない自動運用システムに適用する場合, 正確な訓練用データを自動収集するための機能が強く求められる.

本稿では, 対象とする電子データを電子メールとし, 電子メール特有のドメイン情報を基にした訓練用データの自動収集手法を提案する. 提案手法では, 電子メールの機密性定義をドメイン情報に基づいて実施し, 機密, 非機密それぞれの訓練用データを過去に送受信した電子メールの中から自動的に収集, 選定する. これにより, 高精度な機密メール検出システムの自動運用が可能となる.

提案手法のアプリケーション適用例として, 機密メール検出機能を備えた電子メールアーカイブが挙げられる. 電子メールの保管を目的としたアーカイブシステムに対して機密情報検出技術を適用することにより, 故意や誤りによる機密メールの外部送信が検出可能となり, 情報漏洩事故発生時の事後追跡効率を向上させることができる. 人手を介さない自動運用が求められる電子メールアーカイブにおいて, 提案手法に基づく訓練用データの自動収集機能は機密メール検出機能を実現するための鍵となる. 本稿では, 正規表現・学習型フィルタ併用方式に基づいて提案手法を電子メールアーカイブに実装し, 評価によりその有効性を実証する.

本稿の構成は以下のとおりである. まず, 2 節で先行研究について述べる. 3 節では, 訓練用データ自動収集手法を提案する. 4 節では, 提案手法と正規表現・学習型フィルタ併用方式の組み合わせによる, 電子メールアーカイブへの機密メール検出機能実現方式を説明する. 5 節では 4 節の実現方式に対する評価結果を報告する. 最後に, 6 節で本稿のまとめを行う.

2. 先行研究とその課題

機械学習を使用したテキスト分類において, 訓練用データ収集の人手負担を低減させる手法として co-training が良く知られている [4], [5]. これは, まず少数の正しく分類された訓練用データを人手で用意し, それらを学習して得られる分類器により未知データを分類し, 分類確度の高いものを新たに訓練用データとして採用するという処理を繰り返す手法である. 学習結果を元に, 分類済みのデータを訓練用データとして再利用していると解釈することも可能である. 初めに訓練用データを人手収集すれば, 以降は自動的に学習が実施され分類器の精度が向上していくという点で, co-training は自動運用に適した手法であると言える.

しかしながら, システムの導入時に少数とはいえデータ収集を人手で実施するというのは, 機密情報検出において大きな課題となる. 当然ながら, 機密情報検出では機密情報を多数取り扱うため, 一部の限られた管理者を除く関係者は, 機械学習に必要な機密データの閲覧権限を持たないのが普通である. このことは, 訓練用データ収集の困難さに直結する. そのため, 一般の技術者による機密情報検出システムの構築・導入が極めて困難となる.

また, co-training ではシステムの使用環境変化などに伴い, 扱う機密情報に変化があった際, それらを正しく検出することができない. そのため, 定期的に人手による訓練用データの再収集が必要となり, システム自動運用への妨げとなる.

3. 訓練用データ自動収集手法

本節では, 機密メール検出における訓練用データ自動収集手法を提案する. なお, 以下では機密メール検出の実施を想定している環境を自組織と表現し, その他の組織を外部組織と呼ぶ.

3.1 電子メールの機密性定義

訓練用データの自動収集を実現するためには, 電子メールが機密メールか否かを形式的に定義可能である必要がある. そこで, 電子メール特有の情報であるドメイン情報を利用して, 電子メールの機密性を定義することを考える. なお, ここではメールアドレスにおける “@”以降の任意のピリオドの次の文字からメールアドレス末尾までの文字列を総称してドメインと呼ぶ. 例えば, メールアドレス “xxx@abc.def.com”に対しては

- “abc.def.com”
- “def.com”

などがドメインに該当する. ドメイン情報を基に, 電子メールの機密性を次のように定義する.

定義 3.1. 適当なドメインからなる集合 $\mathcal{D}$ と, メールアドレスが記述される適当なヘッダフィールドからなる集合 $\mathcal{F}$ をそれぞれ 1 つずつ選んで固定しておく. このとき, ある電子メール M について, 集合 $\mathcal{F}$ に属するヘッダフィールドに記述されたメールアドレスが全て集合 $\mathcal{D}$ に属するいずれかのドメインを持つとき, M を機密と定義し, $domain(M) = 1$ と表記する. 電子メール M が機密でないとき, M を非機密であると定義し, $domain(M) = 0$ と表記する.

集合 $\mathcal{D}$ は機密と判定したいドメインの集合に対応する. 例えば, 自組織のドメイン 1 つのみを $\mathcal{D}$ の要素と定義すれば, 自組織内でやり取りされた電子メールのみが機密と定義され, 外部組織とやり取りされた電子メールは全て非機密として扱われる. また, $\mathcal{D}$ として自組織と複数の関連組織のドメインを設定すれば, 設定された関連組織とやり取りされた電子メールを機密に含めることが可能である.

集合 $\mathcal{F}$ は, 機密判定の際にどのメールアドレスを使用するかを指定する役割を果たす. 通常, 考えられる $\mathcal{F}$ の定義は

$$\mathcal{F} = \{\text{From},\ \text{To},\ \text{Cc}\} \cup (\text{Bcc に対応したヘッダフィールド集合}) \qquad (1)$$

である. 要件に応じて Bcc や Cc を集合 $\mathcal{F}$ から取り除くことも可能である.

式 (1) で集合 $\mathcal{F}$ を定義する場合, 電子メールの宛先, 発信元

のいずれか一方には必ず自組織ドメインが含まれる．そのため，集合 $\mathcal{D}$ には自組織ドメインを含める必要がある．自組織ドメインを集合 $\mathcal{D}$ に含めない場合，定義 3.1 により，全ての電子メール M に対して $domain(M) = 0$ となり，意味のある結果は得られない．このように多くの場合，集合 $\mathcal{D}$ には自組織ドメインを含めるのが自然である．

3.2 訓練用データ候補抽出

訓練用データは過去に送受信された電子メールの中から収集する．収集した電子メールの中から，定義 3.1 に基づき，機密，非機密用の訓練用データとして使用する電子メールの候補をそれぞれ決定する．以下ではこの工程を候補抽出と呼ぶ．

まず，単純な候補抽出方式を示す．

候補抽出方式 3.2. 各電子メール M に対し，次の規則に従って候補抽出を実行する．

- $domain(M) = 1$ のとき，M を機密用候補とする．
- $domain(M) = 0$ のとき，M を非機密用候補とする．

候補抽出方式 3.2 を採用することで，過去に送受信された電子メールの中から機密用，非機密用の訓練用データ候補を自動抽出することができる．

候補抽出方式 3.2 には問題点が 1 つ存在する．それは，例えば $domain(M) = 0$ となるような電子メール M の中には，過去に誤って外部送信された機密メールが含まれている可能性がある，ということである．候補抽出方式 3.2 では，このような電子メールを非機密用候補として抽出し，結果として正しい機械学習が実施されない．このように，例外的にではあるが，ドメイン情報による機密性定義と実際のメール内容から判断される機密性とは必ずしも一致しないことがある．そこで，この点を改良することを考える．

改良の基本方針は，候補抽出の際にドメイン情報のみを使用するのではなく，別の情報を併せて利用するというものである．訓練用データ候補は過去に送受信された電子メールから抽出するため，既に機密メール検出システムを導入している環境下において，これらの電子メールに対する機密検出が過去に一度実施されているという状況が想定可能である．そこで，全ての訓練用データ候補は過去に機密検出が実施されていると仮定し，この検出結果を利用することを考える．ここで，記号を定義する．

定義 3.3. 機密情報検出を既に実施済みの電子メール M について，M が機密と判定されているとき $result(M) = 1$ と表記し，M が非機密と判定されているとき $result(M) = 0$ と表記する．

過去の検出結果を利用した候補抽出の改良方式を示す．

候補抽出方式 3.4. 各電子メール M に対し，次の規則に従って候補抽出を実行する．

- $domain(M) = result(M) = 1$ のとき，M を機密用候補とする．
- $domain(M) = result(M) = 0$ のとき，M を非機密用候補とする．
- $domain(M) \neq result(M)$ のとき，M を機密用候補としても非機密用候補としても抽出しない．

候補抽出方式 3.4 では，定義 3.1 により非機密と定義された電子メールであっても，過去の検出で機密と判定されていれば訓練用データ候補として抽出されることはない．従って，過去に誤って外部送信したような電子メールを誤って機械学習することが起こらないと期待できる．

候補抽出方式 3.4 の考え方は，未知データを分類結果に基づいて訓練用データとして採用するという意味で co-training に共通する部分がある．しかしながら，提案手法は，ユーザによって指定された電子メールのドメイン情報を訓練用データ候補抽出の基礎としている点で co-training と大きく異なる．例えば，環境変化により電子メールの内容に変化が生じた場合であっても，機密と判定したいドメイン情報に変更がなければ，機械学習はその変化に対応し，自動的に検出条件は修正される．また，機密と判定したいドメイン情報に変更が生じた場合には，設定したドメインの追加・削除のみを行えば良い．これは，検出条件を再生成する場合も，既存の検出条件の更新を続ける場合も同様である．ドメイン情報の変更はユーザ自身で容易に行えるため，訓練用データを再収集する場合と比較して人手負担を大幅に低減することが可能である．

3.3 訓練用データ選定

前項で述べた候補抽出方式 3.4 により，過去に送受信された電子メールの中から訓練用データ候補が自動抽出される．しかしながら，ここで抽出された候補を全て訓練用データとして使用することが最良とは限らない．例えば次のような場合，訓練用データを選定する必要性が生じる．

- 抽出された全候補データに対して機械学習を実施するために必要な時間が，システムが許容する時間を越える場合
- 抽出された機密用，非機密用の候補データに大きな件数差がある場合

1 つ目の例は，連続運転しているようなシステムにおいて，機械学習に割当可能な時間が限られているような状況が当てはまる．この場合，予めシステムに設定された上限値以下となるように訓練用データを選定する必要がある．

一方，2 つ目の例はシステム要件に依る制約ではなく，機械学習に基づいたテキスト分類手法が持つ性質に依るものである．一般に，機械学習に基づいたテキスト分類では，分類させたいクラス毎に訓練用データを準備し機械学習を実施する．この際，クラス間で訓練用データ件数にバラつきがあると機械学習の精度は低下することが知られており，学習アルゴリズムの工夫，訓練用データの再標本化 (resampling) など，様々な観点から精度低下を防止するための研究が行われている [6], [7]．ここでは特に訓練用データの再標本化に注目する．訓練用データの再標本化方法には，大きく分けて，少数クラスのデータ件数を増加させる oversampling と，多数クラスのデータ件数を減少させる undersampling の 2 種類がある．この undersampling が訓練用データの選定に相当する．訓練用データ件数のバラつきを低減させるようなデータ選定により検出精度が向上すること

は, 文献 [3] においても検証済みである.

本研究では, 検出対象が電子メールであり, 訓練用データは過去に送受信された電子メールの中から収集されるため, 訓練用データとして使用可能なデータ件数を十分に入手可能な状況を想定している. そこで, undersampling の考え方に基づき, 上述の 2 点を同時に解決することが可能な訓練用データの選定方式を以下に提案する.

選定方式 3.5. 機密用, 非機用の訓練用データ件数合計の上限値が u 件と定められているとする. このとき, 訓練用データの候補抽出により抽出された機密用候補 n_1 件, および非機密用候補 n_0 件を, 次の規則によりそれぞれ $\widetilde{n}_1$ 件, $\widetilde{n}_0$ 件に選定する.

- $n_1 + n_0 \leq u$ のとき, $\widetilde{n}_1 := n_1$, $\widetilde{n}_0 := n_0$ とする.
- $n_1 + n_0 > u$ かつ $n_1 < u/2$ のとき, $\widetilde{n}_1 := n_1$, $\widetilde{n}_0 := u - n_1$ とする.
- $n_1 + n_0 > u$ かつ $n_0 < u/2$ のとき, $\widetilde{n}_1 := u - n_0$, $\widetilde{n}_0 := n_0$ とする.
- $n_1 + n_0 > u$ かつ $n_1 \geq u/2$ かつ $n_0 \geq u/2$ のとき, $\widetilde{n}_1 := u/2$, $\widetilde{n}_0 := u/2$ とする.

選定方式 3.5 は, 機密用, 非機密用として選定される訓練用データ件数の合計が上限値 u に最も近くなるようにしながら, 件数差を最小とするような方式となっている. 訓練用データ候補が上限値 u に対して十分に存在するときは, 機密, 非機密が同件数 $u/2$ となるように訓練用データが選定される. 機密用, 非機密用の候補データ件数に差があり同件数ずつ選定できない場合においても, 件数差が最小となるように訓練用データが選定される.

以上が提案する訓練用データの自動収集手法である. 本節で述べた訓練用データの自動収集手法の流れを図 1 に示す.

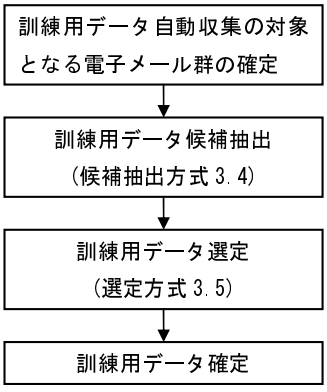

図 1 訓練用データ自動収集の流れ

4. 電子メールアーカイブ適用

本節では, 前節で提案した訓練用データ自動収集手法と正規表現・学習型フィルタ併用方式を組み合わせ, 電子メールアーカイブに対して自動運用可能な機密メール検出機能を実現する方式を述べる.

4.1 正規表現・学習型フィルタ併用方式

電子メールアーカイブへの適用方式を述べる前に, ここで簡単に正規表現・学習型フィルタ併用方式の説明をしておく. 詳しくは文献 [2], [3] を参照されたい. 以下では, 正規表現・学習型フィルタ併用方式のことを単に併用方式とも呼ぶ.

正規表現・学習型フィルタ併用方式とは, 予め設定した機密用語を文字列照合により検出する正規表現フィルタと, 機械学習を利用した学習型フィルタとを併用した機密情報の検出方式である. 全ての入力文書に対して, 正規表現フィルタと学習型フィルタの両方による機密判定が行われ, 少なくともどちらか一方のフィルタが機密と判定した文書を総合的に機密と判定する.

併用方式では, 正規表現フィルタと学習型フィルタが役割を分担することにより, 簡易な条件設定で高精度な検出を実現可能としている. 正規表現フィルタでは, 「社外秘」などの機密等級や「システム開発計画書」といった機密文書名など, 業務規則等で規定された形式的な機密用語のみを検出条件として設定し, 定型的な機密文書の検出に専念する. このような検出条件であれば, 検出対象文書の内容に精通していなくとも条件設定が形式的に可能であり, 作成された検出条件は作成者に大きく依存しない. 一方, 学習型フィルタでは, 正規表現フィルタにより検出困難であるような文書を検出する. 例えば, 使用環境により変動しやすい機密用語や, 業務内容に関する専門的な知識を必要とする機密用語は人手による設定が難しいため, 学習型フィルタによって検出する.

併用方式により, 学習型フィルタの初期検出精度の低さ, 環境変化への弱さが正規表現フィルタによって補完される. 逆に, 正規表現フィルタにおける検出条件設定の困難さが, 学習型フィルタを併用することで緩和される. このように, 併用方式は正規表現フィルタと学習型フィルタの両者の特長を活かしながら, 弱点を補完し合うような方式となっている.

4.2 機密メール検出機能の実現方式

電子メールアーカイブにおける機密メール検出機能の実現方式を述べる. 機密メール検出で使用する検出フィルタには併用方式を採用し, 学習型フィルタが使用する訓練用データについては提案手法を用いて収集する. これにより, 機密メール検出機能を備えた電子メールアーカイブの自動運用を実現する. 電子メールアーカイブの構成を図 2 に示す.

電子メールアーカイブは, 予めシステム管理者等により設定された 1 時間などの周期毎に蓄積処理を起動し, 1 周期間内に新規に送受信された電子メールを対象に処理を実施するものとする. この前提の下, 電子メールアーカイブにおける機密メール検出機能の起動タイミングと動作を図 3 のように設計する. 機密メール検出機能は蓄積処理の直後に起動され, 1 周期間内に送受信された電子メールに対して機密検出を実施する. この際, 正規表現フィルタを使用し, 定義 3.1 に基づくドメイン判定を同時に実施しておく. 機密検出が終了すると, 検出対象であった電子メール群はそのまま訓練用データ自動収集の対象となる. 提案手法を用いて訓練用データを収集した後, 機械学習を実施して学習型フィルタの検出条件を更新する.

以上のように機密メール検出機能を動作させることで, 提案

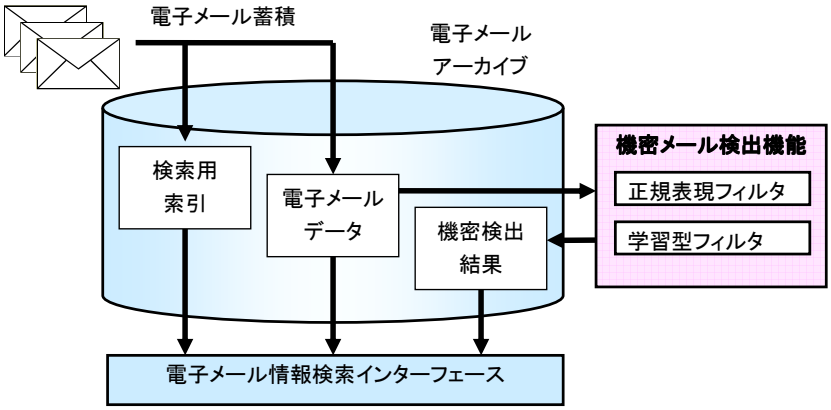

図 2　電子メールアーカイブの構成

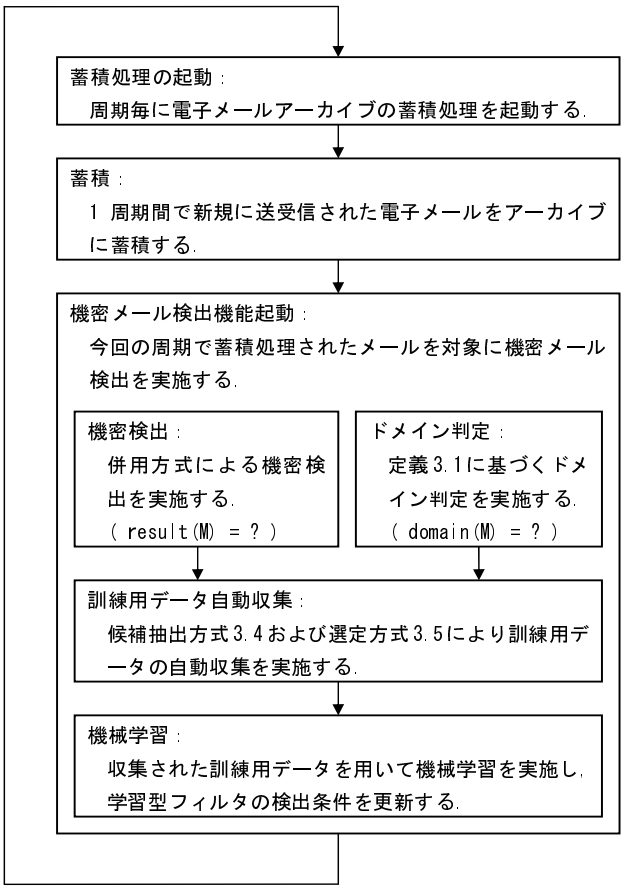

図 3　機密メール検出機能の起動タイミングと動作

する訓練用データ自動収集手法を活用することができる．訓練用データを収集する段階では既に機密検出が実施済みであるため, 候補抽出方式 3.4 および選定方式 3.5 による訓練用データ自動収集が可能となる．

4.3 学　習　率

訓練用データ選定方式 3.5 では, 機密用, 非機用の訓練用データ件数合計の上限値 u が設定されている必要がある．電子メールアーカイブ適用においては, この上限値 u を学習率というパラメータによって決定する．

定義 4.1.　訓練用データの候補抽出によって抽出された全データに対して, 実際に訓練用データとして選定するデータ件数の割合を学習率と呼ぶ.

定義 4.1 により, 学習率 α $(0 < \alpha \leq 1)$ を与えたとき, 訓練用データの候補抽出方式 3.4 により抽出された機密用候補 n_1 件, 非機密用候補 n_0 件に対して, 上限値 u が

$$u = \alpha(n_0 + n_1) \tag{2}$$

と定義される．すなわち, 電子メールアーカイブ適用においては, 上限値 u を定数として予め設定するのではなく, 学習率 α により, 機密メール検出機能の起動毎に式 (2) を用いて自動算出する．

学習率を導入する利点として, 機密メール検出機能の起動周期を変更したときに, 上限値 u を都度変更する必要がないということが挙げられる．起動周期が変更されれば, 当然, 1 度に処理される電子メール件数や機械学習に割当可能な時間は変化するため, 適切な上限値 u もまた変化する．学習率を使用することで, 適切な上限値 u が自動的に決定される．

また, 学習率を使用することにより, 1 日を通してシステムが機械学習に割り当てる時間をほぼ一定とすることができる．電子メールアーカイブの蓄積処理は 1 時間などの比較的短い周期で起動されることが多いため, 1 周期間内に新規に送受信される電子メール件数には周期毎に大きな差がある．例えば, 昼間は電子メール件数が多く, 夜間は少ないなどである．一方で, 1 日単位などの長い周期から見れば, その間に送受信される電子メール件数は環境によってある程度定まっている．従って, 学習率を使用することにより, 日単位で見れば選定される訓練用データ件数をほぼ一定とすることができる．結果として, 1 日を通してシステムが機械学習に割り当てる時間はほぼ一定となる．更に, 1 日を通して送受信される電子メールから, 偏りなく訓練用データを選定することも可能となる．

4.4 初期学習運用と本運用

前項までで示した機密メール検出機能の実現方式には, 1 つ解決すべき問題がある．それは, 機密メール検出機能の初期導入時は機械学習が全く為されていないため, 学習型フィルタによる

機密検出が行えないということである. 従って, 訓練用データ収集においても, 機密検出結果を利用する候補抽出方式 3.4 が使用できず, 結果として訓練用データの自動収集を実施できない.

そこで, この問題点に対応するために, システム運用の段階を初期学習運用と本運用の 2 段階に分ける.

定義 4.2. 初期学習運用とは, 正規表現フィルタのみによる機密検出を行い, 学習型フィルタによる検出を実施しない運用形態であると定義する. 一方, 本運用とは, 正規表現・学習型フィルタ併用方式による機密検出を行う運用形態であると定義する. 訓練用データ収集と機械学習は, 初期学習運用, 本運用の両方において同様に実施される.

本運用は通常の運用形態である. 一方, 初期学習運用は, 機械学習をある程度進め, 学習型フィルタの検出条件を実用レベルにするための準備期間と捉えることができる. 機密メール検出機能の初期導入時は初期学習運用を実施し, ある程度学習が進めば本運用に切り替える. 但し, 初期学習運用期間中も正規表現フィルタによる検出は実施されるため, 機密メール検出の能力が失われているわけではない. 初期学習運用と本運用による運用の流れを図 4 に示す.

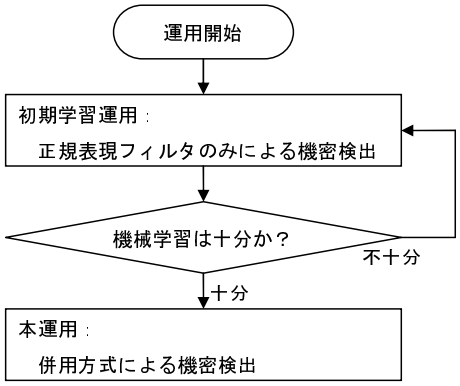

図 4 初期学習運用と本運用

以上により, 電子メールアーカイブにおける機密メール検出機能が実現され, システムの初期導入時より人手の訓練用データ収集が不要な自動運用が実現される.

5. 評　　価

本節では, 前節で述べた電子メールアーカイブへの適用方式を実装し, 評価した結果を報告する.

5.1 評価内容・条件

評価の基本方針は, 前節までで示した電子メールアーカイブにおける機密メール検出機能を, 実運用に近い形態で動作させ, 検出精度を調べるというものである. 検出精度は機密メール検出機能の起動周期ごとに算出し, 検出精度の遷移を中心に調べる. 特に, 初期学習運用期間を適切に設定することで, 機械学習が自動で進み, 本運用に切り替えたときに検出精度が向上することを確認する.

評価実験は以下の条件で実施する.

- **評価データ**

使用する評価データは 2007 年から 2008 年の 2 年間に送受信された業務メール約 65,000 件である. 評価データにおける機密性の正解は定義 3.1 に基づくドメイン判定によって定義する. 但し, 集合 $\mathcal{D}$ は

$$\mathcal{D} = \{ \text{自組織のドメイン} \} \cup (\text{関係会社のドメイン計 110 個}) \quad (3)$$

によって定義し, 集合 $\mathcal{F}$ は式 (1) を採用する. この条件下での評価データの内訳を表 1 に示す.

表 1 評価データ内訳

機密メール	53,992 件
非機密メール	11,655 件
計	65,647 件

- **起動周期**

本評価では 2 年分の電子メールを月毎の 24 区間に分割し, 1ヶ月周期で機密メール検出機能を起動する. 機密メール検出機能は, 図 3 に示した手順に従い, 機密検出, 訓練用データ収集, 機械学習を実施する. 検出精度は起動周期毎に算出する.

起動周期が 1ヶ月というのは, 実際の運用と比較するとかなり長い周期である. これは, 実運用における電子メール処理件数と比較し, 評価データの件数が少ないことに起因する. 本評価では, 1 周期当たりの平均電子メール処理件数が約 2,700 件となり, これは, 1 人 1 日当たり 20 件の電子メールを送受信するような 3,000 人規模の組織において, 機密メール検出機能を 1 時間周期で起動させるときの処理件数に相当する.

- **初期学習運用期間・学習率**

初めの 4 周期分, すなわち 4ヶ月分を初期学習運用期間として割り当て, 残りの 20ヶ月分を本運用により動作させる. また, 学習率は 20%とする. すなわち, 初期学習運用期間中に, およそ 2,000 件のデータを学習させるような状況を想定している.

- **正規表現フィルタ設定条件**

本評価で使用する正規表現フィルタの検出条件における, 設定用語の内訳を表 2 に示す. 但し, 検出条件は正規表現により記述され, 単純な単語の羅列とはならないため, 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 正規表現フィルタの検出条件

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

- **評価指標**

評価指標として, 再現率 (recall), および適合率 (precision) を使用する. 本評価において, 各指標は次式で定義される.

$$\text{再現率} = \frac{\text{正しく機密と判定されたメール件数}}{\text{機密メール件数}} \quad (4)$$

$$\text{適合率} = \frac{\text{正しく機密と判定されたメール件数}}{\text{機密と判定されたメール件数}} \quad (5)$$

これらの指標は, テキスト分類, 情報検索において一般的に用いられる指標である [8]. 再現率は検出漏れの少なさを, 適合率は過剰検出の少なさをそれぞれ表しており, 共に数値が高いほど精度が良い. なお, 機密情報検出においては検出漏れの低減が最優先されるため, 再現率の向上が最も重要となる [2].

5.2 評価結果

評価結果を図 5, 6 に示す. 図はそれぞれ再現率, 適合率に対応し, 各図には併用方式による検出精度と正規表現フィルタのみの検出精度を記している. また, 横軸は機密メール検出機能の起動回数を表している.

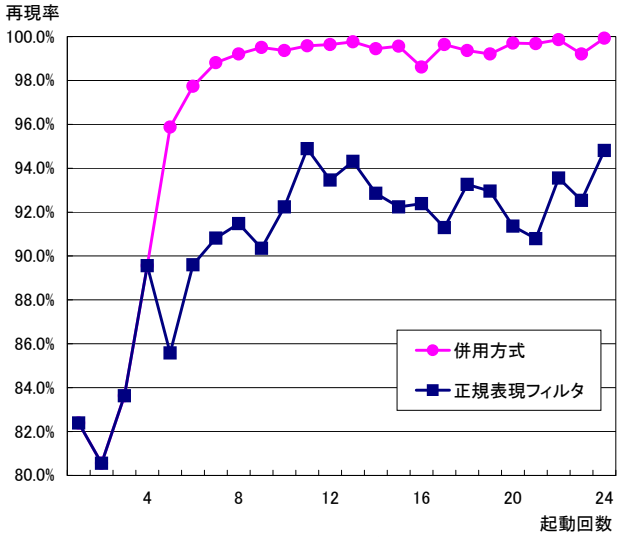

図 5 再 現 率

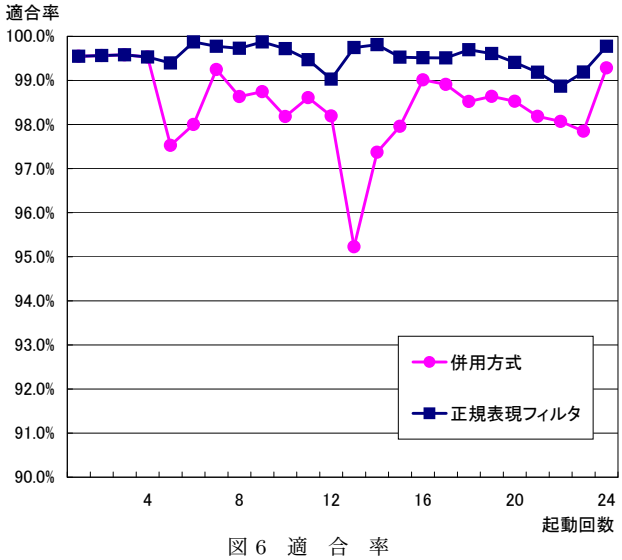

図 6 適 合 率

正規表現フィルタの検出精度は, 初期学習運用を続けたときの検出精度に相当する. 図 5 より, 本運用への移行に伴い再現率が大きく向上していることが確認できる. このことから, 初期学習運用期間に実施された 2,000 件程度のデータ学習により十分な精度となることが分かり, 更に, 提案した訓練用データ自動収集手法が有効に機能していることが分かる. 前項で述べたように, 実際の運用では 1 日で数万件程度の電子メールが蓄積されることが想定されるため, 長くとも 1 日の初期学習運用期間を経る事で, 本運用に移行可能と考えられる.

ここで, 本運用期間中に実施された全データに対する検出精度を表 3 にまとめる.

表 3 検証項目 1 本運用期間の検出精度

フィルタ	再現率	適合率
併用方式	99.2%	98.2%
正規表現フィルタ	92.1%	99.5%
学習型フィルタ	98.7%	98.4%

併用方式において再現率 99%以上の結果が出ており, 十分に精度が出ていると言える. 適合率については, 併用方式の原理から必然的に学習型フィルタを併用することで精度が低下するが, 本運用以降後も 98%と高い精度を保っていることが分かる.

5.3 分 析

図 6 によると, 13 回目の機密検出において, 過剰検出が他の回と比較して非常に多いことが分かる. この原因を特定するため, この期間に過剰検出された電子メールを調査した. 評価データにおいて, 上記期間は 2008 年 1 月に対応するのであるが, 調査の結果, 過剰検出件数 124 件中 77 件が, 2008 年 1 月より加入したメーリングリストによるものであることが判明した. 以後, このメーリングリストをメーリングリスト A と呼ぶ.

過去にやり取りされていなかったような電子メールが新たに加わると, 当然, 学習型フィルタの検出精度は低下する. 上述の結果はこの事実を反映している. ところで, 図 6 より, 2008 年 1 月以降は大きな検出精度低下は起きていない. 例えば, 2008 年 4 月にはメーリングリスト A による電子メールが約 200 件受信されているのであるが, 対応する 16 回目の処理において過剰検出はあまり起きていない事が確認できる. この事実より, 人手による調整なしに学習結果が自動的に反映され, 検出精度を向上させていると分析することができる.

以上の分析内容を比較実験により検証する. 実験内容は次のとおりである.

実験内容. 適合率が著しく低下した 13 回目の機密検出以後, 機械学習の実施を止め, その後の検出精度遷移を調べる. これにより, 前項までの結果と併せて, 学習型フィルタの検出条件更新有無による精度比較を行う.

分析が正しければ, 学習型フィルタの検出条件更新を止めたことにより, 13 回目以降の機密検出において検出精度が低下する. 特に, 13 回目以降の他の回においても, メーリングリスト A が検出精度低下要因として現れると予想される.

比較実験の結果を図 7, 8 に示す. 図はそれぞれ再現率, 適合率に対応し, 各図には 12 回目以降の機密検出結果の遷移を記している. 図中の “併用方式”は, 図 5, 6 の値と同一であり, “併用方式 (更新なし)”が比較実験による検出精度を表している. 実験の結果, 再現率に関しては学習型フィルタの検出条件更新有無による大きな変動はないが, 適合率に関して, 検出条件の更新

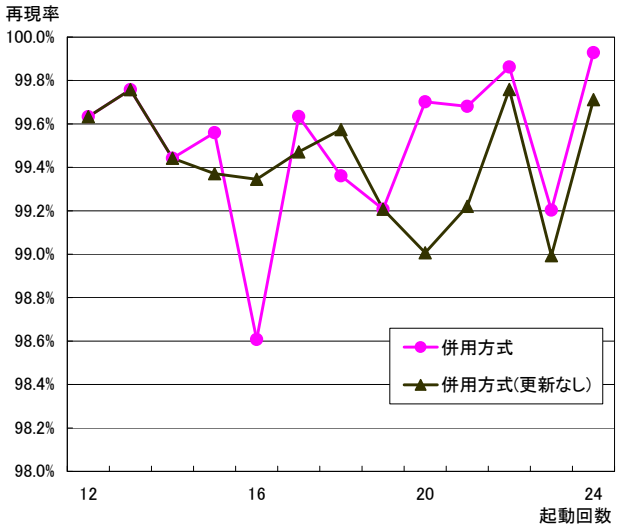

図 7　検出条件更新有無による再現率比較

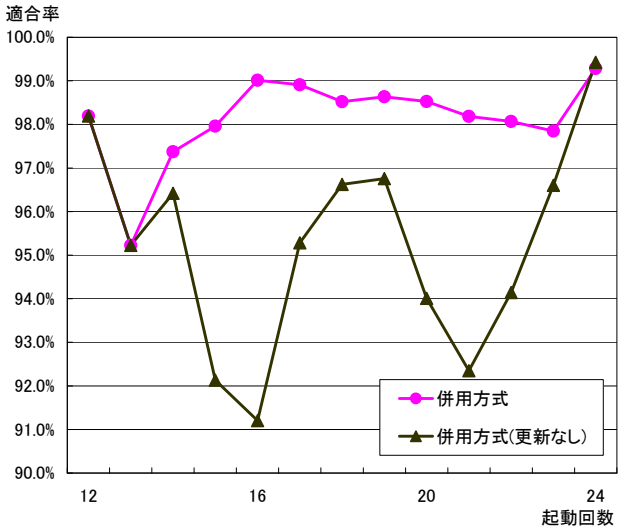

図 8　検出条件更新有無による適合率比較

を止めた事により大幅に検出精度が低下している．この結果をメーリングリスト A と関連付けるため，新たな指標 X を

$$X = 1 - \frac{\text{メーリングリスト } A \text{ による過剰検出件数}}{\text{機密と判定されたメール件数}} \qquad (6)$$

で導入する．ここで，式 (5) より適合率が

$$\text{適合率} = 1 - \frac{\text{過剰検出件数}}{\text{機密と判定されたメール件数}} \qquad (7)$$

と書き換えられることに注意すると，常に

$$\text{適合率} \leq X \leq 1 \qquad (8)$$

が成立する．指標 X は，適合率低下においてメーリングリスト A が与える影響度と解釈することができる．$X = 1$ であれば，メーリングリスト A による過剰検出はなく，逆に，$X =$ 適合率であれば過剰検出は全てメーリングリスト A によるものである．

指標 X と適合率の比較を図 9 に示す．図 9 は，図 8 と対応しており，学習型フィルタの検出条件を更新しない場合の適合率，および，対応する指標 X の値を示している．図 9 より，メーリングリスト A が 13 回目以降の機密検出においても検出精度低

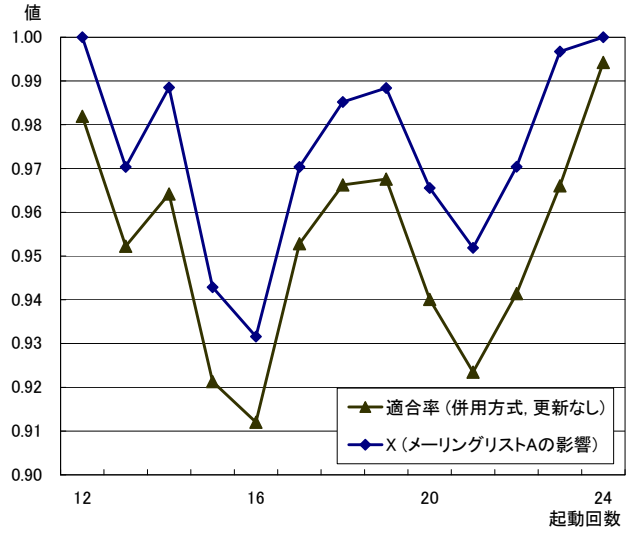

図 9　メーリングリスト A が適合率低下に与える影響

下の要因として大きく現れていることが分かる．

以上より，分析結果の妥当性が確認でき，結論として，提案手法の有用性が実証された．

6. おわりに

本稿では，機械学習を利用した機密メール検出システムにおける訓練用データ自動収集手法を提案した．また，提案手法と正規表現・学習型フィルタ併用方式を組み合わせることで，電子メールアーカイブに対して自動運用可能な機密メール検出機能を実現する方式を提案した．これらの実装，評価を通じて，提案手法を用いることで，高い検出精度を維持したまま機密メール検出システムの自動運用が実現可能であることを確認した．

今後は，機密メール検出システムの異なる自動運用形態や，電子メール以外のデータを対象としたシステムの研究開発を進めていく予定である．

文　献

[1] 日本ネットワークセキュリティ協会, “2008 年情報セキュリティインシデントに関する調査報告書”, http://www.jnsa.org/

[2] 加藤守, 柴田秀哉, 郡光則, “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講演論文集 (2), pp. 157-158, 2009.

[3] 柴田秀哉, 加藤守, 郡光則, “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”, 第 8 回情報科学技術フォーラム講演論文集 (2), pp. 159-160, 2009.

[4] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training”, In Proceedings of the 11th Annual Conference on Computational Learning Theory, pp.92-100, 1998.

[5] S. Kiritchenko and S. Matwin, “Email Classification with Co-Training”, In Proc. of the 2001 Conference of the Centre for Advanced Studies on Collaborative Research, 2001.

[6] B. Zadrozny, J. Langford and N. Abe, “Cost-sensitive learning by cost-proportionate example weighting”, ICDM 2003, pp. 435- 442.

[7] A. Liu, J. Ghosh and C. Martin, “Generative Oversampling for Mining Imbalanced Datasets”, DMIN 2007, pp. 66-72.

[8] D. Lewis, “Evaluating Text Categorization”, In Proceedings of the Speech and Natural Language Workshop, pp. 312-318, 1991.