

ベイジアンフィルタと社会ネットワーク手法を統合した迷惑メールフィルタリングとその最適統合法

大 福 泰 樹[†] 松 浦 幹 太[†]

近年、電子メールの普及にともない迷惑メールが急増し、社会問題となっている。本稿では、ベイズ理論を用いて統計的に迷惑メールをフィルタリングするベイジアンフィルタと、メールの送受信関係から抽出される社会ネットワークを用いてメールアドレスのホワイトリスト・ブラックリストを作成する手法とを統合し、その統合手法の実験・評価を行う。また、その統合法について Ham 判定重視やベイジアンフィルタによる SNA の補正などの改良手法を提案し、その評価を行う。

Integration of Bayesian Filter and Social Network Technique for Combating Spam E-Mails Better

HIROKI OHFUKU[†] and KANTA MATSUURA[†]

In recent years, spamming has been increasing rapidly with the spread of E-mails and has become a social problem. In this paper, we integrate two anti-spam techniques; one is a statistical method for filtering spam, "Bayesian filter", and the other is an e-mail address listing method using "social network analysis (SNA)" which exploits the sender-recipient relationship. We evaluate the integrated spam filtering method. Then we propose some effective improvements such as biasing for ham classification and modification of SNA by Bayesian filter, and evaluate them.

1. はじめに

インターネットの世界では、古くからスパムメールやジャンクメールなどと呼ばれる迷惑メールが存在する。商業的な広告宣伝、勧誘などのダイレクトメールをはじめとして、政治や宗教的宣伝のメール、いたずらや嫌がらせのメール、不幸の手紙のようなチェーンメール、非法なビジネスへの勧誘や情報の提供など、受信者の意思にかかわらず、一方的に繰り返し送りつけられるメールがそれにあたる。

迷惑メールは、近年のインターネット社会の発展とともに大きな問題となっている。何万通ものメールを、非常に少ないコストで一度に送信できることから、大量に広告メールを送信する悪質な業者が増加しているためである。2001年にはインターネット上でやりとりされる電子メール全体の8%だった迷惑メールが、2003年には全体の50%を上まわり、2004年には全体の65%に達したとの調査報告も存在する¹⁾。最近ではフィッシングメールと呼ばれる詐欺メールの問題も浮

かび上がっている。それに対抗して様々な迷惑メール対策技術が考案され使用されているが、迷惑メール送信者による送信方法も巧妙化しており、なかなか十分な対策が行えないというのが現状である。このような状況下で迷惑メールを撲滅すべく様々な角度から対策が研究されている。

まずあげられるのは、本文やヘッダなどメール内容に含まれる情報をもとにフィルタリングを行うという、迷惑メールに対する最も単純な対策方式である。従来、メール内容によって迷惑メールをフィルタリングする方法はルールベースのものが中心であったが、Graham の論文²⁾以降、ベイジアンフィルタリングという統計的な手法が注目を集めている。Robinson の論文³⁾では、Graham の方式に基づいて新しいベイジアンフィルタが提案されており、それは SpamBayes などいくつかのフリーの迷惑メールフィルタや、市販のアンチスパムソフトなどに実装されている。

それ以外に、ホワイトリストを使用する方式^{4)~7)}も考案されている。これは、受信者が持つホワイトリストに登録されている送信者からのメールのみを、受信者に提示するという方式である。ホワイトリストに登録されていない送信者に対しては登録手続きが指示さ

[†] 東京大学生産技術研究所

Institute of Industrial Science, The University of Tokyo

れる．そして、その登録手続きを行ったユーザのみがホワイトリストに登録され、以後メールの送信を許可される．また、そうした登録手続きに頼らずに、メールの送受信関係から社会ネットワークを抽出し、ホワイトリストを自動生成する手法⁸⁾もある．

また、最近フィッシング対策として徐々に注目を集めつつある送信者認証技術や、法律面からの対策などもある．

本稿では、ベイジアンフィルタと社会ネットワーク手法⁸⁾を統合し、迷惑メールを効率的にフィルタリングする手法を提案し、その評価を行う．それぞれの特徴を端的に述べれば、ベイジアンフィルタは広い範囲のメールをカバーできるが誤りが比較的多く、社会ネットワーク手法によるフィルタリングは一部のメールしかカバーできないが誤りはほぼ0といってよいほど少ない．これら2つを統合することにより、両者のメリットをうまく引き出すとともに互いの欠点を補い合い、それぞれのフィルタを単独で用いた場合よりも判定精度の高いフィルタを実現したい．しかし、不適切な手法で統合すれば2つのフィルタがむしろ互いに悪影響を及ぼし、単独のフィルタよりも判定精度が落ちてしまう恐れもある．本稿では、最適な統合法を目指していくつかの統合法を提案し、それぞれでフィルタを単独で用いた場合よりも判定精度が高くなることを示すとともに、異なる統合法を比較考察する．そのためにまず2章でベイジアンフィルタについて、3章で社会ネットワーク手法 (SNA: Social Network Analysis) について説明し、4章でそれらの統合法について述べる．5章で今回行った実験について説明し、6章で結果と評価について述べる．

2. ベイジアンフィルタ

ベイジアンフィルタでは、まず過去に受信した迷惑メール (Spam) と正当なメール (Ham) のデータをもとにして、ある単語 w を含む電子メールが迷惑メールである確率 $p(w)$ を計算する．そしてこの $p(w)$ を用いて判定対象の電子メール m が迷惑メールである確率 $p(m)$ を計算し、その確率がある一定の閾値 t を上回ったものを迷惑メールと判断する．

英語のメールの場合は単語がスペースで区切られているが、日本語のメールでは一部に句読点がある以外単語に区切りがないので、形態素解析が必要になることもある．本稿の実験では、簡単のため、英語メールのみを扱う．

ベイジアンフィルタには、判定したメール中の単語を新たに学習し、出現確率のデータを更新できるとい

う特徴がある．それにより、以後の迷惑メールの判定精度が向上していく．たとえば、迷惑メール業者が送信するメールの内容が変化するとともに、フィルタで遮断する迷惑メールの基準も変化する．また、それぞれのユーザが受信する迷惑メールや正当なメールの傾向に合わせて、フィルタリング基準も変化していく．

2.1 Paul Graham 方式

ここでは、Graham^{2),9)}によって提案された、ベイズ理論を用いた迷惑メール確率計算の方法について、説明を行う．まず、過去に受信した迷惑メールと正当なメールに含まれている単語の頻度情報の学習データがあるとする．ある単語 w に対して、 w を含むメールが迷惑メールである確率 $p(w)$ は、学習データを用いて以下のような計算で求める．

$$p(w) = \frac{\frac{b(w)}{n_{bad}}}{a \cdot \left(\frac{g(w)}{n_{good}} \right) + \frac{b(w)}{n_{bad}}} \quad (1)$$

$g(w)$ 正当な電子メールにおける単語 w の頻度

$b(w)$ 迷惑メールにおける単語 w の頻度

n_{good} 正当な電子メール数

n_{bad} 迷惑メール数

a バイアス (定数)

文献 2) では、 $a = 2$ としている．これは、正当な電子メールを誤って遮断してしまうこと (false positive) の方が、迷惑メールがフィルタを通過してしまうこと (false negative) よりも損害が大きいという考えから、正当な電子メールの誤遮断が起こりにくいようにバイアスをかけるためとされている．

新しいメールが届くと、それを単語に分解し、最も特徴的な M 個の単語 ($w_1, \dots, w_M$) を抽出する．ここで特徴的というのは、その単語の迷惑メール確率が 0.5 から遠く離れていることとする．文献 2) では、 $M = 15$ としている．そのメール m が迷惑メールである確率 $p(m)$ は、 $p(w_1), \dots, p(w_M)$ の統合確率で表すことができ、次のように計算される．

$$p(m) = \frac{p(w_1) \cdots p(w_M)}{p(w_1) \cdots p(w_M) + (1 - p(w_1)) \cdots (1 - p(w_M))} \quad (2)$$

そして、 $p(m)$ がある閾値 t を上回った場合、そのメールは迷惑メールと判定される．文献 2) では、 $t = 0.9$ としている．閾値を 0.5 ではなく 0.9 と高めに設定している理由は、false positive を避ける方向にバイアスをかけるためである．

2.2 Gary Robinson 方式

Graham 方式をもとにして、Gary Robinson が提

案した方式³⁾である. この Robinson 方式では, 単語ごとの迷惑メール確率 $f(w)$ を以下のように求める.

まず, Graham 方式の単語ごとの迷惑メール確率 $p(w)$ を, バイアスをかけずに求める.

$$p(w) = \frac{b(w)}{b(w) + g(w)} \quad (3)$$

その $p(w)$ を用いて, $f(w)$ は次のように計算される.

$$f(w) = \frac{s \cdot x + n \cdot p(w)}{s + n} \quad (4)$$

ここで, x は今まで 1 度もメール中に出現していない単語が初めてメールに出現したときに, そのメールが迷惑メールである予測確率とし, s (strength) をその予測に与える強さとする. また n は単語 w の出現回数とする. x と s の値は, フィルタのパフォーマンスが最適化されるように設定すべきであるが, とりあえずは, $x = 0.5$, $s = 1$ が妥当であるとされている.

Graham 方式と比較してこの方式が優れているのは, 単語 w の出現回数が少ない場合 ($n = 0$ を含めて) をうまく扱える点である. たとえば, Graham 方式ではある単語 w がスパムメールのみに数回出現した場合, そのメールの迷惑メール確率 $p(w)$ は 1 になってしまうが, その程度の情報で単語 w に最大の迷惑メール確率を与えてしまうのはやりすぎであろう. しかし一方, Robinson 方式では, 単語 w の総出現回数 n が小さい場合には $p(w)$ の比重が小さくなるようにできているので, まだ情報不足であるということを $f(w)$ に暗に加味することができる. そして学習が進むに従い, 総出現回数 n が大きくなってゆき, $f(w)$ の値は漸近的に $p(w)$ の値に近づいてゆく. また, $n = 0$ の場合には $f(w) = x$ となる.

さらに, あるメールが迷惑メールである確率は次の I で与えられる.

$$H = C^{-1} \left(-2 \ln \prod_w f(w), 2n \right) \quad (5)$$

$$S = C^{-1} \left(-2 \ln \prod_w (1 - f(w)), 2n \right) \quad (6)$$

$$I = \frac{1 + H - S}{2} \quad (7)$$

C^{-1} は逆 χ^2 関数 (inverse chi-square function) を意味する. H は Hamminess (ノンスパム性), S は Spamminess (スパム性) の略で, I はそれらを統合した指標 (Indicator) である.

最終的に迷惑メールかどうかを判定する閾値 t については特に指定はないが, 判定結果を迷惑メールと正当なメールに 2 分するのではなく, I が 0.5 に近い場

合はどちらともいえない (Unsure) と判定することにより, 誤判定を減らすことができるとしている. 今回の実験では閾値 t は 0.5 とし, I が 0.5 に一致した場合は, 「疑わしきは罰せず」という方針に基づき Ham と判定することにした.

3. 社会ネットワークを利用したフィルタリング手法

メールの送受信関係の社会ネットワーク分析 (SNA: Social Network Analysis) によって, メールアドレスのホワイトリスト・ブラックリストを構築する手法である. 複数のユーザのメールデータからメールアドレスの送受信関係ネットワークを構成する方法^{10)~12)}もあるが, 本稿では単独のユーザが自分自身で受信したメールのみでフィルタリングができることを目指しているため, 今回は単独のユーザが受信したメールから送受信ネットワークを構築し迷惑メールを判別するという, 社会ネットワーク手法⁸⁾に注目する.

この手法では, ユーザの受信したメールの From, To, Cc ヘッダに注目し, メールアドレスの社会的ネットワークを構築する. まず, ユーザのメールボックスのメールのヘッダに現れるすべてのアドレスに対応したノードを作る. そして, From アドレスのノード (図 1 の例では, Alice) から, 同じメールヘッダに現れる他のすべてのアドレスのノード (図 1 の例では, Bob, Charlie, David, Ed) へ枝を張って接続する. この枝は, 両端のノード間に送受信関係があることを意味する. そうしてできたネットワークから, ユーザ自身のアドレスを除くことにより, ユーザの周りのメールアドレスネットワークが構築される. このネットワークには複数の独立したメールアドレスのネットワークと孤立したノードが含まれているが, この独立したメールアドレスネットワークの 1 つ 1 つのことをコンポー

```
From: "Alice" <alice@example.com>
To: "Bob" <bob@example.com>,
    "Charlie" <charlie@example.com>
Cc: "David" <david@example.com>,
    "Ed" <edward@example.com>
```

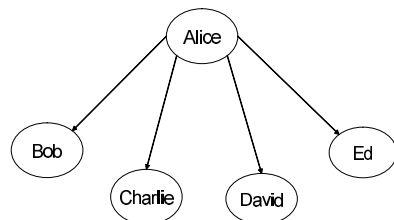


図 1 送受信関係の例

Fig. 1 Example of sender-recipient relation.

ネットと呼ぶことにする。これらのコンポーネントが信頼できるアドレスのネットワーク（Ham コンポーネント）に相当するのか、もしくは迷惑メール関連のアドレスのネットワーク（Spam コンポーネント）に相当するのかを判別するために、ネットワークの親密度、つまりクラスティング係数を計算する。ユーザの知人同士は互いに知り合いである可能性が高く、その間でメールのやりとりがあることは十分考えることができ、信頼できるアドレスのコンポーネントのクラスティング係数は高くなるが、迷惑メールの被害者同士が互いに知り合いであることはほとんどないので、迷惑メールに対応するコンポーネントのクラスティング係数は非常に低くなる（0 になることがほとんどである）。この手法では、クラスティング係数 C が 0.01 より小さい場合、そのコンポーネントを迷惑メール関連のアドレスからなるコンポーネントと見なし、そこに含まれるアドレスをブラックリストに加える。 C が 0.1 より大きい場合には、そのコンポーネントを信頼できるアドレスのコンポーネントと見なし、そこに含まれるアドレスをホワイトリストに加える。また C が 0.01 と 0.1 の間になったときは、そのコンポーネントに含まれるアドレスに対する判断は保留し、ブラックリストにもホワイトリストにも加えない。ここで問題となるのは、どの程度の大きさのコンポーネントならクラスティング係数による判定対象と見なすかということである。なぜなら、信頼できるアドレスのコンポーネントであっても、そこに含まれるノードがまだ少ない場合には、クラスティング係数が非常に小さくなることが考えられるからである。その場合には、信頼できるアドレスのコンポーネント、迷惑メール関連のアドレスのコンポーネントと見なししてしまうことになる。よって、そのコンポーネントを判定対象に加えるかどうかを決めるコンポーネントサイズ（ノード数）の閾値 t_n を設定する必要がある。本稿の実験では $t_n = 10$ と設定した。

この手法をあるメールデータに適用した実験⁸⁾によると、ブラックリスト、ホワイトリストには分類誤りは 1 つもなかったが、メールアドレスのうち 50% の判断が保留された。判定精度は非常に高く有効な方法であるが、一部のメールしか判定できない手法であり、広い範囲のメールをカバーできるフィルタと併用すべき手法であるといえるだろう。

3.1 クラスティング係数

コンポーネントに含まれるノード間の親密度を表す指標となるクラスティング係数は、以下のように計算される。

あるノード i の周りに k_i 個の隣接ノードが存在するとすると、隣接ノード間には最高で $\frac{k_i(k_i-1)}{2}$ のコネクションが存在する。実際の隣接ノードどうしのコネクションの数を E_i とすると、ノードあたりのクラスティング係数 C_{node} は

$$C_{node} = \frac{2E_i}{k_i(k_i - 1)} \quad (8)$$

のように表せる。そして、コンポーネントに含まれるすべてのノード（ただし $k_i > 1$ のもの）について C_{node} を計算し、その平均をもってそのコンポーネントのクラスティング係数とする。

$$C = \frac{1}{N_2} \sum_i \frac{2E_i}{k_i(k_i - 1)} \quad (9)$$

ただし、 N_2 は少なくとも 2 つ以上の隣接ノードを持つノードの個数を表す。

3.2 判定・学習プロセス

メールの送受信関係をもとに構築されたブラックリスト・ホワイトリストを利用して、日々受信するメールを逐次的に分類する方法は次のようになる。

- (1) コンポーネントサイズが t_n 以上の各コンポーネントのクラスティング係数を計算する。
- (2) クラスティング係数が 0.01 より小さい場合、そのコンポーネントに含まれるアドレスをブラックリストに加え、0.1 より大きい場合はホワイトリストに加える。
- (3) 判定対象となるメールの送信者アドレスをブラックリストおよびホワイトリストと照合する。
- (4) そのアドレスがブラックリストにあった場合は迷惑メールと見なし、ホワイトリストにあった場合は正当なメールと見なす。リストになかった場合は判定不能とする。
- (5) そのメールの送受信関係を学習する。つまり、送受信関係ネットワークに含まれる適切なコンポーネントにその送受信関係を接続する、もしくは新たなコンポーネントを作成する。
- (6) 以下、(1)～(5)を繰り返す。

4. 迷惑メール対策手法の統合

Robinson 方式のベイジアンフィルタと、社会ネットワーク手法を統合することを考える。それぞれ、ベイジアンフィルタは広い範囲のメールをカバーできるが誤りが比較的多く、社会ネットワーク手法は一部のメールしかカバーできないが誤りはほぼ 0 といっているほど少ない、という特徴がある。これら 2 つをうまく統合することにより、互いの欠点を補い合い、また相乗効果が生まれるようにしたい。図 2 にその統合

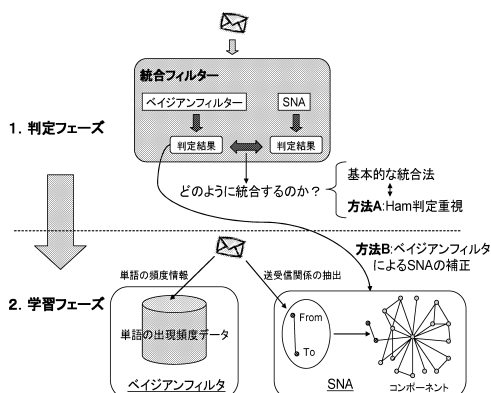


図 2 統合イメージ

Fig. 2 Integration image.

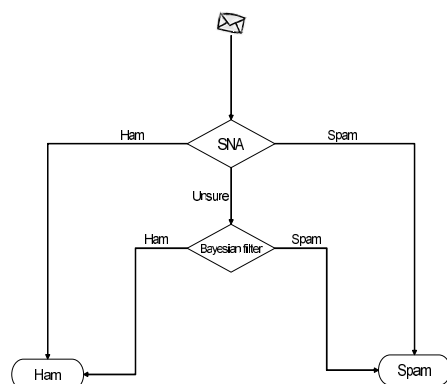


図 3 基本的な統合法

Fig. 3 Trivial method.

イメージを示す。

4.1 基本的な統合法

ベジアンフィルタと社会ネットワーク手法の最も基本的な統合法はどのようなものか。社会ネットワーク手法が、メールアドレスのホワイトリストとブラックリストを作成し、判定精度がベジアンフィルタよりかなり高い手法であることを考えれば、最も当たり前の統合法は図 3 のようになる。つまり、

- (1) まず判定対象となるメールの送信者アドレスが、SNA によるリストに含まれる場合は、ホワイトリストに含まれるなら正当なメール、ブラックリストなら迷惑メールと見なす。
- (2) SNA によるリストに送信者アドレスが含まれていない場合、ベジアンフィルタの判定結果を判定対象のメールに適用する。

以上の方法を基本的な統合法と見なし、4.2 節のような工夫を加えた統合法と比較していくことにする。

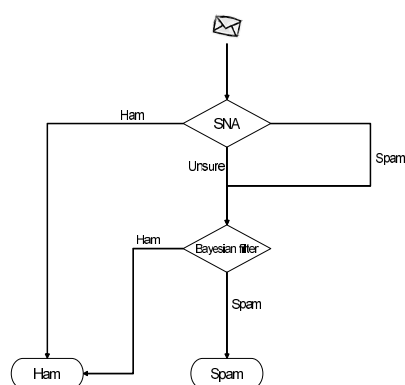


図 4 方法 A : Ham 判定重視

Fig. 4 Method A: Make much of Ham classification result.

4.2 提案手法

ベジアンフィルタと社会ネットワーク手法をうまく統合することにより、互いの欠点を補い合い、また相乗効果が生まれるようにしたい。そのために次のような方法を用いた。

(方法 A) ベジアンフィルタか SNA の少なくともどちらか一方で正当なメールと見なされたら、正当なメールと見なす (Ham 判定重視)。

(方法 B) あるメールがベジアンフィルタで迷惑メールと判定されたとき、学習フェーズにおいてその送受信関係は SNA の Ham コンポーネントには加えない。また逆に、ベジアンフィルタで正当なメールと見なされたとき、その送受信関係は SNA の Spam コンポーネントには加えない (ベジアンフィルタによる SNA の補正)。

方法 A の処理の流れを図 4 に示す。方法 A によって、片方のフィルタが誤遮断 (正当なメールを迷惑メールとして遮断してしまうこと: false positive) をしてしまっても、もう片方のフィルタの判定が正しければその誤りを補うことができ、false positive を減らすことができる。

方法 B の処理の流れを図 5 に示す。方法 B は、社会ネットワーク手法によるホワイトリスト・ブラックリスト作成をベジアンフィルタにより補正するということである。社会ネットワーク手法の欠点としては、迷惑メール送信者が故意、もしくは偶然に受信者の知り合いにメールを同報していた場合、その迷惑メール送信者のアドレスが SNA のホワイトリストのコンポーネントに加えられてしまうということがあるが、方法 B によりそれを防ぐことができる。

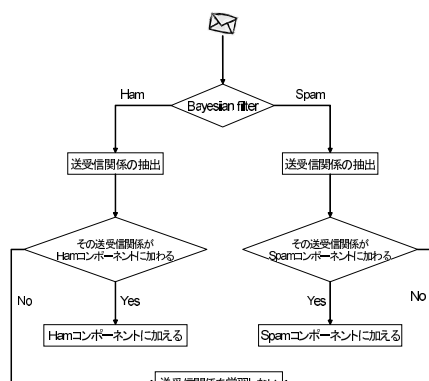


図 5 方法 B: ベイジアンフィルタによる SNA の補正

Fig. 5 Method B: Modification of SNA by Bayesian filter.

5. 実験

実験対象として、2004 年 4 月から 2005 年 3 月までの 1 年間に研究室のある特定の個人が受信した英語の電子メールを使用した。対象となった電子メールは 2,838 通（正当なメール 1,445 通，迷惑メール 1,393 通）であった。このメールデータに対し，時間順に 1 通ずつ，判定とそこに含まれる単語や送受信関係の学習を行った。判定は 4 章のそれぞれの統合手法によって行う。比較のため，統合前の各手法単独の場合についても実験する。

すべてのメールの判定が終わったところで，誤遮断率（FPR: false-positive rate = 正当な電子メールが迷惑メールとして遮断されてしまう確率），誤通過率（FNR: false-negative rate = 迷惑メールが正当なメールと見なされフィルタを通過してしまう確率），誤判定率（ER: error rate = 迷惑メールであるかどうかを誤判定されてしまう確率，つまり誤遮断もしくは誤通過してしまう確率）や，ホワイトリスト，ブラックリストの作成精度などを調べた。

6. 結果と評価

6.1 ベイジアンフィルタのみの場合

はじめに，ベイジアンフィルタ単独の場合の実験結果を表 1 に示す。すべてのメールが迷惑メールと正当なメールに二分され，誤判定も起こった。FPR が比較的高いことが分かる。

6.2 社会ネットワーク手法のみの場合

次に，SNA 単独の場合の判定結果を表 2 に，リスト作成結果を表 3 に示す。判定を SNA 単独で行うという意味では文献 8) と同じであるが，リスト作成に方法 B で補正を加えるか否かで生じる違いを比較し

表 1 ベイジアンフィルタによる判定結果

Table 1 Classification results by Bayesian filter.

成功率	FNR	FPR	ER
97.43% (2,765/2,838)	0.72% (10/1,393)	4.36% (63/1,445)	2.57% (73/2,838)

ており，単なる追実験ではなく，最適統合法を目指すための独自の予備実験となっている。

まず表 2 を見ると，方法 B でリスト作成を補正するかどうかによらず，SNA 本来の特徴が現れている。すなわち，50%あまりが判定不能となるが，ER はきわめて低い。しかし，表 3 のホワイトリスト，ブラックリストの作成精度を見てみると，方法 B を使った場合は誤分類は 0 であったが，方法 B を使わない場合，ホワイトリストに Spam 関連のアドレスが 359 個も含まれてしまった。これは，あるメールの送受信関係によって Spam コンポーネントと Ham コンポーネントが結合されてしまったためと考えられるが，方法 B によりこれらの結合を防ぐことができたということである。今回のデータでは表 2 を見れば分かるように判定精度にあまり差が出なかったが，表 3 に現れているリスト作成精度の差は，今後さらにメールを受信していった場合，迷惑メール判定に大きな差をもたらすと予測できる。

6.3 2つを統合した場合

最後に，2 つのフィルタを統合した場合の結果を，表 4 に示す。方法 A，方法 B を使うかどうかにより，条件 1~4 の 4 通りの統合方法を試した。

まず分かるのは，ベイジアンフィルタ単独の場合よりも，条件 1~4 のすべての統合方法について判定精度が高くなっているということである。たとえば，ベイジアンフィルタ単独の場合の false positive の数は表 1 によると 63 通であったが，表 4 を見ると false positive が 1 番多い条件 2 の場合でも 41 通となっている。これにより，複数のフィルタを統合することが有効であるといえる。

また，SNA 単独の場合，判定可能だったメールは全体のうち半分ほどであったが，統合によりすべてのメールを判定することができた。

さらに，条件 1 と条件 2，条件 3 と条件 4 を比べれば，方法 B の有効性を見ることができる。しかし，方法 B を使うことによって，非常に小さな幅ではあるが FNR は下がり，FPR は上がっており，表 4 から方法 B の有効性をはっきり示すデータは得られなかった。ただし，表 3 からは方法 B によってリスト作成精度がはっきりと上がることが示されており，今後ともメールを受信し続けた場合や，判定対象となるメー

表 2 SNA による判定結果 ($t_n = 10$)
Table 2 Classification results by SNA.

	成功率	FNR	FPR	ER	不明
方法 B なし	46.62% (1,323/2,838)	0.22% (3/1,393)	0.97% (14/1,445)	0.60% (17/2,838)	52.78% (1,498/2,838)
方法 B あり	46.65% (1,324/2,838)	0% (0/1,393)	1.04% (15/1,445)	0.53% (15/2,838)	52.82% (1,498/2,838)

表 3 SNA によるリスト作成結果 ($t_n = 10$)
Table 3 Lists made by SNA.

	Whitelist		Blacklist		Gray	
	Ham	Spam	Ham	Spam	Ham	Spam
方法 B なし	502	359	0	51	12	804
方法 B あり	508	0	0	403	15	811

表 4 統合フィルタによる判定結果
Table 4 Classification results by integrated filters

	方法 A	方法 B	成功率	FNR	FPR	ER
条件 1	×	×	98.17% (2,786/2,838)	0.93% (13/1,393)	2.70% (39/1,445)	1.83% (52/2,838)
条件 2	×	○	98.20% (2,787/2,838)	0.72% (10/1,393)	2.84% (41/1,445)	1.80% (51/2,838)
条件 3	○	×	98.34% (2,791/2,838)	0.93% (13/1,393)	2.35% (34/1,445)	1.66% (47/2,838)
条件 4	○	○	98.41% (2,793/2,838)	0.72% (10/1,393)	2.42% (35/1,445)	1.59% (45/2,838)

ルコーパスを変えた場合などで判定精度にも差が出てくる可能性がある。

またさらに、条件 1 と条件 3、条件 2 と条件 4 を比べれば方法 A の有効性を見ることができるが、これについては両方とも FNR は変わらず、FPR が低下していることが見てとれる。4.2 節で述べたように、方法 A の目的は正当なメールを迷惑メールとして遮断してしまう false positive を減らすことであったが、これは成功したといえるだろう。

条件 1~4 すべてを比較した場合でも、FNR、FPR に多少の上下はあるが、判定誤りを総合的に見た ER については、低い方から順に条件 4、3、2、1 となっている。条件 2、3 と条件 4 を比べれば、方法 A または方法 B の片方だけを使うよりも両方を使ったほうがよいということが分かる。

また、計算コストについてだが、ベイジアンフィルタを単独で用いた場合と 2 つのフィルタを統合した場合とで、処理時間にほとんど差はなかった。ベイジアンフィルタで行われる計算は、文章からの単語の抽出、出現回数のカウント、スコアの算出であり、社会ネットワーク手法で行われる計算は、メールヘッダからの送受信関係の抽出、コンポーネントのクラスタリング係数の算出である。統合による計算コストの増加は無

視してもよい程度のものだったといえる。

7. おわりに

本稿では、統計的に迷惑メールをフィルタリングするベイジアンフィルタと社会ネットワーク手法とを統合し、迷惑メールをよりの確にフィルタリングする手法を提案し、その有効性を示すことができた。

今回、統合により判定精度が向上することを示したが、その結果にどの程度の意味があるのか、他研究との比較を行うことは難しい。なぜなら、そもそも実験に用いたデータが異なる（メールフィルタリングの分野で信頼できる標準的な判定対象データセットは今のところ存在しない）し、実装方法（ヘッダを含めるかどうか、学習方法など）も異なるからである。市販のフィルタと比べることもあまり意味がない。市販のものは、ベイジアンフィルタなどの基幹となる技術の上に、細かいルールを人手で実装していくことにより判定精度を高めたものである。本稿で実装したシステムでも、誤判定されたメールについてその原因を調べ、それを避けるような細かいルールを人手で設定していけば、判定精度を向上させることは可能であろう。本稿の目的は、フィルタの基幹となる技術について改良を加えたり、それら基幹技術の統合を行ったりするこ

とによって、いかに判定精度向上が見込めるかを明らかにすることである。実用的な迷惑メールフィルタが基幹技術に細かいルールを付加したものとしても、基幹技術の判定精度がより高くなれば、より安定した判定精度を持つフィルタとなりうるだろう。

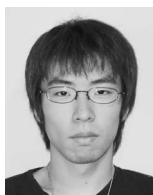
今後は、今回実験に使用したメールデータ以外の様々なデータに対しても判定実験を行い、この統合法の有効性を示す必要があるだろう。また、その際に、ユーザがなんらかのコミュニティに属しているかどうか、属しているならどのようなコミュニティに属しているのかなどによって、SNA によるアドレスネットワークの構成のされ方が異なると予測される。そういったメールデータの持つ要素が判定結果に及ぼす影響などについても、調べてゆきたい。

参 考 文 献

- 1) Symantec. <http://www.symantec.com/>
- 2) Graham, P.: A Plan for Spam (2002). <http://paulgraham.com/spam.html>
- 3) Robinson, G.: A Statistical Approach to the Spam Problem, *Linux Journal*, Vol.107 (2003).
- 4) Gabber, E., Jakobsson, M., Matias, Y. and Mayer, A.: Curbing Junk E-Mail via Secure Classification, *Financial Cryptography '98*, LNCS 1465, pp.198–213, Springer (1998).
- 5) Mailblocks. <http://www.mailblocks.com/>
- 6) Jakobsson, M., Linn, J. and Algesheimer, J.: How to Protect Against a Militant Spammer, Cryptology ePrint archive, report 2003/07 (2003).
- 7) Hall, R.J.: Channels: Avoiding Unwanted Electronic Mail, *1996 DIMACS Symposium on Network Threats*, pp.85–103, American Mathematical Society (1997).
- 8) Boykin, P.O. and Roychowdhury, V.: Leveraging Social Networks to Fight Spam, *IEEE Computer*, Vol.38, No.4, pp.61–68 (2005).
- 9) Graham, P.: Better Bayesian Filtering, *2003 Spam Conference* (2003).
- 10) Tyler, J.R., Wilkinson, D. and Huberman, B.A.: Email as Spectroscopy: Automated Discovery of Community Structure within Organizations, preprint. <http://xxx.lanl.gov/abs/cond-mat/0303264>
- 11) Ebel, H., Mielsch, L.-I. and Bornholdt, S.: Scale-Free Topology of E-Mail Networks, *Physical Rev. E*, Vol.66, No.035103 (2002).
- 12) Caldarelli, G., Coccetti, F. and Rios, P.D.L.: Preferential Exchange: Strengthening Connections in Complex Networks, *Physical Rev. E*, Vol.70, No.027102 (2004).

(平成 17 年 11 月 25 日受付)

(平成 18 年 6 月 1 日採録)



大福 泰樹

昭和 57 年生。平成 16 年 3 月東京大学工学部電子情報工学科卒業。平成 18 年 3 月東京大学大学院情報理工学系研究科電子情報学専攻修士課程修了。ネットワークセキュリティ、特に迷惑メール対策に興味を持つ。



松浦 幹太 (正会員)

昭和 44 年生。平成 9 年 3 月東京大学大学院工学系研究科電子工学専攻博士課程修了。同年 4 月東京大学生産技術研究所助手。平成 10 年 4 月同講師。平成 12 年 4 月東京大学大学院情報学環講師(生産技術研究所兼任)。平成 14 年 4 月東京大学大学院情報学環助教授(生産技術研究所兼任)。平成 16 年 4 月東京大学生産技術研究所助教授。情報セキュリティ、リスク管理等の研究に従事。博士(工学)。著書に『情報セキュリティ概論』(共著, 昭晃堂, 1999)等。電子情報通信学会, IEEE, ACM 等各会員。日本セキュリティ・マネジメント学会非常任理事。