情報処理学会論文誌 Vol.54 No.2 1002–1011 (Feb. 2013)

推薦論文

ベイジアン方式と機械学習の併用による スパムメールフィルタリング

山口 博之[1,†1] 角 朝香[1,†2] 杉井 学[2,a)] 松野 浩嗣[1,b)]

受付日 2012年3月20日, 採録日 2012年11月2日

概要：近年のスパムメールの増大は，世界中の電子メール利用者を悩ませている．スパムメール対策として様々なメールフィルタの開発が行われているが，その中で現在最も広く利用されているのが，ベイズ理論に基づいてスパムメールを分類するベイジアンである．bsfilter は，ベイジアンフィルタの 1 つであるが，多くのプラットフォームに対応できることなどから，利用者が増えている．しかしこのシステムは，正規メールの正解率は高いが，スパムメールの正解率が低いという欠点がある．そこで，本論文では先に我々が開発した機械学習を応用したメールフィルタリング手法と bsfilter を組み合わせ，さらに bsfilter のメール判定に用いられる閾値の変更を行うことで正規メールとスパムメールともに，正解率の高いメールフィルタリング手法を提案する．

キーワード：スパムメール，ベイジアンフィルタ，機械学習，併用フィルタ

Spam Mail Filtering System with the Combinational Use of Bayesian and Machine Learning Methods

Hiroyuki Yamaguchi[1,†1] Sayaka Kado[1,†2] Manabu Sugii[2,a)]
Hiroshi Matsuno[1,b)]

Received: March 20, 2012, Accepted: November 2, 2012

Abstract: Recently, the increase of spam mails leads us to serious problems all over the world. Various mail filters had been developed for spam mails, and the most common filtering system is based on the baysian theory. Bsfilter is one of the bayesian filters and it has been popularly used because it can be applied to many platforms. Bsfilter can classify regular mails with high accuracy rate but it can not classify spam mails with such a high accuracy rate. In this paper, we propose a method for a mail filtering system that combines bsfilter and machine learning system BONSAI. The combinational performance of these two systems is investigated by examining the combination order of bsfilter and BONSAI and by changing the threshold of bsfilter.

Keywords: spam mail, bayesian filter, machine learning, combinational filter

1. はじめに

電子メールはインターネットや携帯電話網の普及により多くの人に利用されている．しかし，受信者にとって必要でないメール，いわゆるスパムメールが電子メール利用者を悩ませている．スパムメールは電子メールの 1 度に大量

1 山口大学大学院理工学研究科
Graduate School of Science and Engineering, Yamaguchi University, Yamaguchi 753–8512, Japan
2 山口大学大学情報機構メディア基盤センター
Madia and Information Technology Center, Yamaguchi University, Ube, Yamaguchi 755–8611, Japan
†1 現在，株式会社日立システムズ
Presently with Hitachi Systems, Ltd.
†2 現在，日本ラッド株式会社
Presently with Nippon RAD Inc.
a) manabu@yamaguchi-u.ac.jp
b) matsuno@sci.yamaguchi-u.ac.jp

本論文の内容は 2010 年 10 月の電気・情報関連学会中国支部連合大会にて報告され，情報処理学会中国支部長により情報処理学会論文誌ジャーナルへの掲載が推薦された論文である．

に送信できるという特徴などを悪用して無差別に送りつけられるが [1]，その問題点は，受信者にとって必要でないメールが大量に送られてくることが原因で，必要なメールを見落としてしまう危険性などにある．スパムメールの送信手法は様々存在するが，送信者情報を偽ったメールや知人からのメールを装ったケースなどが増えてきており，手口は年々巧妙化してきている [2]．

スパムメールに対抗する手段として，スパムメールと必要な正規メールを自動的に分類する様々なスパムメールフィルタリングが開発されている．その中でも分類精度が高く，現在主流のフィルタシステムは，1998 年に Sahami ら [3] が発表したベイズの定理に則ったものが多い．そして，ベイズの定理を応用したスパムメールフィルタは，Graham によってその有効性が確認され [4]，いくつかの改良のアイデアが組み込まれた方式 [5] が多くのシステムの基礎になっている．

ベイズの定理に基づいてスパムメールを分類するベイジアンフィルタの 1 つに bsfilter [6] がある．bsfilter は正規メールの正解率は非常に高いが，スパムメールの正解率が低いという特徴を持つ．なぜなら，正規メールの誤判定を少なくするために，多くの判定の難しいスパムメールを正規メールへと分類しているからである．現在，スパムメールの流通量は正規メールのそれよりはるかに多く，スパムメールの誤判定による正規メールへの混入をより少なくすることは，メール利用者にとって非常に重要である．bsfilter の場合，スパムメールの誤判定を少なくするためには分類判定に用いられる閾値を変更し，通常正規メールと誤判定されるメールをスパムメールと判定されるようにすればよい．しかし，それだけでは正しく判定できていた正規メールを誤判定する可能性が生じる．なぜなら，閾値付近に分布しているメール群は，その文章特徴に曖昧なものが多く，閾値の小さな変更であっても分類結果が大きく変化してしまうためである．そこで，bsfilter と先に我々が開発した機械学習を応用したメールフィルタリング [7] を組み合わせることで，正規メールの正解率を維持したまま，スパムメールの正解率を上昇させる手法について提案する．さらに，併用フィルタリングシステムを構築し，流通する電子メールの分類を行うことで，本提案手法の有効性を確認する．

2. 機械学習システム BONSAI

機械学習システム BONSAI は，概念学習の 1 つである確率的近似学習（Probably Approximately Correct Learning：PAC 学習）の学習パラダイムに基づき開発されたシステムである [8]．BONSAI は，一次元の記号列データを対象としており，あらかじめ分類しておいた正の例と負の例と呼ばれる 2 つの学習例を与えることで，それらを分類する規則を導き出す．

BONSAI は「インデキシング」と「決定木」により分類規則を作成する．インデキシングは学習例に含まれる要素に対してグルーピングを行い，要素数を減少させる．学習例の特徴を維持したまま要素数の減少ができれば，学習の効率化につながる．また，決定木はインデキシング後の記号列をノードに配置し，正の例と負の例を最も高い確率で分類することができる規則を提示する．

3. BONSAI を応用したメールフィルタリング

我々はこれまでに，**図 1** に示す手順で，BONSAI と Rough Reading [9] の手法を用い，スパムメール群と正規メール群を学習例として BONSAI に入力することで，2 つのメール群を分類する特徴パターンの抽出を試みてきた [7]．Rough Reading とは，ある文章中の出現頻度の高

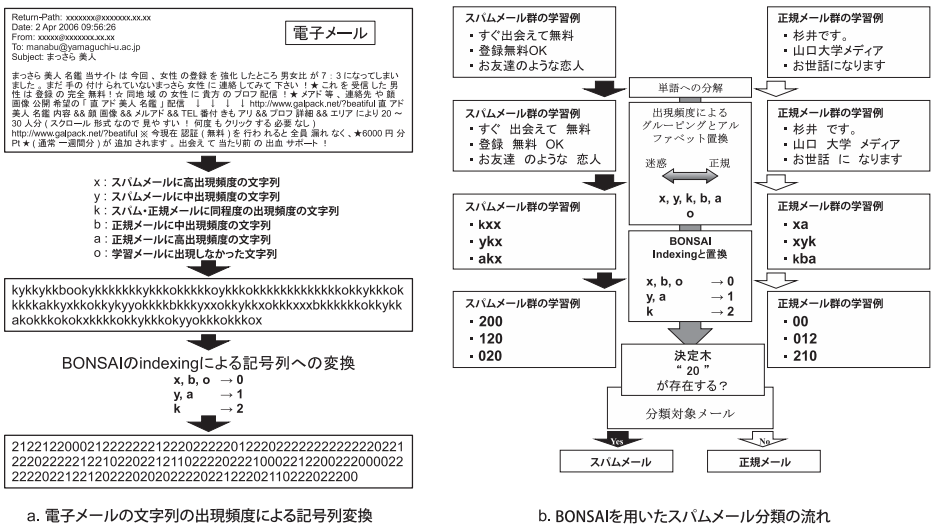


図 1 BONSAI を用いたスパムメール分類の流れ
Fig. 1 Flow of Spam Mail Filtering System by BONSAI.

い単語が文章内容を代表する重要単語ととらえ，その重要単語の文章中の出現パターンによって，その文章の特徴を表す特徴化の方法である．具体的には与えられた文章中の各単語について，以下の方法で出現数や出現頻度を基に重要単語の選定を行い，重要度の度合いに応じた文字に変換し，BONSAI に入力する．BONSAI はインデキシングによってさらにこの文字をグルーピングし，特徴量を減少させたうえで，特徴配列の抽出と決定木を作成してメールの分類を行う．この手法は，言語特有の文章構造の解析を行うことなく，単語の出現頻度とその出現順のパターンのみを利用するため，記述されているメールの言語に依存しないフィルタリングが可能である．ここまでの手順の例を図 1 a に示している．説明を容易にするために，この図は日本語の場合を例にしている．図 1 b では，BONSAI によってスパムメールと正規メールからこれらを分離する決定木を作成するプロセスを示している．詳しくは 3.4 節で説明する．

3.1 単語の抽出

学習用メールとして分類したスパムメール群，正規メール群に含まれる単語の抽出を行う．単語の抽出には漢字かな（ローマ字）変換プログラム KAKASI（2.3.4）[14] を使用した．スパムメール群（正規メール群）から抽出した重複を含むすべての単語数を $E_s(E_h)$ とし，スパムメール群（正規メール群）から抽出した単語 w_i の出現数を $C_s(w_i)(C_h(w_i))$ とする．なお，スパムメール群と正規メール群ではメール中から抽出する単語数が大きく異なる場合が多い．そのため式 (1) を用いて，スパムメール群と正規メール群の抽出単語数を基に補正を行い，スパムメール群，正規メール群での抽出単語 w_i の出現数を再計算する．すなわち，スパムメール群の抽出単語 w_i の出現数 $N_s(w_i)$ は式 (2)，正規メール群の抽出単語 w_i の出現数 $N_h(w_i)$ は式 (3) から算出する．学習メール群全体での抽出単語 w_i の出現数 $N_{all}(w_i)$ は式 (4) のようになる．

$$r = \frac{E_s}{E_h} \tag{1}$$

$$N_s(w_i) = C_s(w_i) \tag{2}$$

$$N_h(w_i) = C_h(w_i) \times r \tag{3}$$

$$N_{all} = N_s(w_i) + N_h(w_i) \tag{4}$$

3.2 単語の出現率と出現頻度の算出

学習例のメール群に含まれる各単語の出現率とスパムメール群および正規メール群での出現の偏りを表す出現頻度をそれぞれ算出する．Graham は，スパムメールおよび正規メール中で出現頻度の高い 15 単語を利用して，分類対象メールのスパム確率を計算する方法を提案した [4]．また，Robinson は，Graham の方式に改良を加え，単語の出現頻度に応じた適切なスパム確率の計算式を提案した [15]．いずれの方式も受信したメール全体から得られる平均的なスパム確率を求めており，たとえ学習例全体での合計出現回数が多い単語であっても，一部のスパムメールや正規メールのみに繰り返し出現する単語は，出現頻度が低くなるように補正する処理が含まれている．我々の提案する方式は，単語の出現頻度を基に，BONSAI を用いた局所的な特徴抽出を行うため，Graham および Robinson の方式と異なり，一部のスパムメールや正規メールにだけに現れる単語であっても，特徴の 1 つとして利用できる方法で単語の出現率および出現頻度の算出を行っている．なお挨拶語や，日本語の場合は助詞なども取り除くことなく，出現率および出現頻度の算出を行っている．これは，記述されているメールの言語に依存しないフィルタリングを可能にするために，言語特有の文章構造の解析を行うことなく，分かち書きされた文節や単語の出現頻度と出現順のパターンのみでメールの特徴を抽出するためである．単語 w_i の出現率 $rate(w_i)$ は式 (5)，スパムメール群での単語 w_i の出現頻度 $p(w_i)$ は式 (6) から算出する．単語の出現回数は，挨拶語や助詞などの頻出単語と，ほとんど使われない専門用語などの名詞では大きく差が開く．そこで，単語の出現率の分布範囲を小さくし，後のアルファベット変換を容易にするために自然対数によって変換した．

$$rate(w_i) = \log(N_{all}(w_i)) \tag{5}$$

$$p(w_i) = \frac{N_s(w_i)}{N_{all}(w_i)} \tag{6}$$

3.3 アルファベット変換

文章の特徴を強く表すパターンの抽出を目的としているため，単語の記号列への変換は学習メール群全体に一定回数以上出現した単語について行う．記号列変換対象とする単語の選別に用いる出現回数基準値は，多くの単語の出現回数が 0 から 1 回に集中していることが分かっているので，全単語の出現回数の平均値や中央値ではなく，最もよく出現する単語の出現率の 1/2 の値を基準に検討した．次のような予備実験を行って，メールの分類精度が最も高かった値を採用した．基準値を決定するための予備実験結果を**表 1** に示す．

予備実験には，6.1 節と同じメールデータセットを使用した．メールの分類精度の比較のための学習数はスパムメール，正規メールともに 100 通を使用した．学習例のメール群に含まれるすべての単語の中で，最もよく出現する単語の出現率を rate_max とし，その 1/2 の値を基準として ±1，±2 の 5 つの値を基準値 rate_base として変化させながら，それぞれの値を採用した場合のメール分類精度の比較を行った．分類対象スパムメール数，分類対象正規メール数はそれぞれ「$test_{spam}$」，「$test_{ham}$」で表す．$correct_{spam}(correct_{ham})$ は「スパムメール（正規メール）

表 1 基準値決定のための分類結果
Table 1 Result of the comparative experiment to determine rate_base.

$trainingmail$	200				
$rate\_base$	$\frac{rate\_max}{2}-2$	$\frac{rate\_max}{2}-1$	$\frac{rate\_max}{2}$	$\frac{rate\_max}{2}+1$	$\frac{rate\_max}{2}+2$
$AccuracyRate_{spam}$	94.7%	97.1%	95.7%	95.1%	95.8%
$AccuracyRate_{ham}$	97.4%	98.1%	93.2%	98.6%	93.8%

表 2 出現頻度による記号列変換表
Table 2 Conversion of word based on an appearance frequency of the word.

変換記号	$p(w_i)$
x	0.8 以上
y	0.8 未満，0.6 以上
k	0.6 未満，0.4 以上
b	0.4 未満，0.2 以上
a	0.2 未満

を正しく分類したメール数」である．$AccuracyRate_{spam}$ は「スパムメールを正しく分類した割合」であり式 (7) から，$AccuracyRate_{ham}$ は「正規メールを正しく分類した割合」であり，式 (8) から算出する．

$$AccuracyRate_{spam} = \frac{correct_{spam}}{test_{spam}} \tag{7}$$

$$AccuracyRate_{ham} = \frac{correct_{ham}}{test_{ham}} \tag{8}$$

表 1 から分かるように，rate_max の 1/2 の値の ±2 を基準値とした場合は，全体的にメールの分類正解率が減少している．この原因としては，記号列変換される単語が過多であるか過少であることが考えられ，±3 以上に設定しても高い正解率の向上は見込まれないと判断し，表 1 の中でも最も分類精度が高い，次の式 (9) を基準値として採用した．

$$rate\_base = \frac{rate\_max}{2} - 1 \tag{9}$$

学習例全体におけるスパムメール群での単語の出現頻度 $p(w_i)$ の値を基に，**表 2** のように「x，y，k，a，b」の 5 つの記号からなるアルファベットに置き換える．「x，y，k，a，b」のいずれにも置き換えられなかった単語，すなわち，出現率が式 (9) の $rate\_base$ 未満の単語は，学習例において重要でない単語として扱い，「o」に置き換える．つまり，式 (6) で示す $p(w_i)$ は，正規メールとスパムメールから抽出された全単語 $N_{all}(w_i)$ のうち，ある単語のスパムメール群での出現数 $N_s(w_i)$ であり，「x，y」に置き換えられた単語はスパムメール群に特徴的に現れる単語，「a，b」に置き換えられた単語は正規メール群に特徴的に現れる単語，「k」に置き換えられた単語はスパムメール群，正規メール群の両メール群に現れる単語を示している．この単語の出現頻度と変換記号の関係を図示すると**図 2** のようになる．

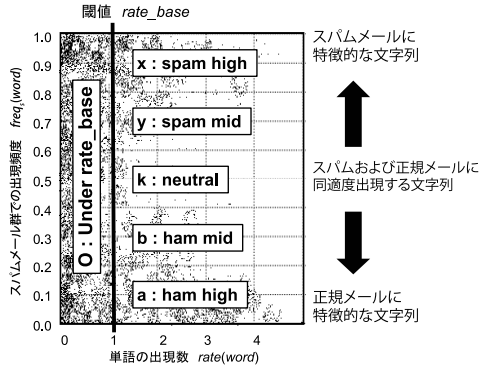


図 2 文字列の出現頻度と変換記号の関係
Fig. 2 The relationship between appearance frequencies and converted symbols of words.

3.4 BONSAI による分類規則の作成

出現頻度による記号列変換によって 6 つの記号からなるアルファベットで表現された学習例のメールを BONSAI に投入し，スパムメール群と正規メール群を分類する規則の生成を行う．つまり BONSAI は，図 1 に示すように，学習例を構成するアルファベットの 6 つの記号をインデキシングし，学習例であるスパムメール群と正規メール群を最も効率良く分類できる規則をインデキシングと決定木の組合せで出力する．学習例が異なれば，BONSAI が設定するインデキシングも異なる．決定木は，出現頻度を表すアルファベット {x，y，a，b，k，o} をインデキシングして得られたインデキシング記号の配列をノードに持ち，各ノードのパターンが電子メール中に存在するか存在しないかによって分類を行う．たとえば図 1 b の場合，電子メールに含まれる文字列の出現頻度によって，電子メールの内容は 6 つの記号からなるアルファベット {x，y，a，b，k，o} にすべて置き換えられる．つまりスパムメールに含まれる “すぐ出会えて無料” という文字列は，“kxx” の記号に変換される．その後，BONSAI が提示したインデキシングによって，アルファベットの 6 つの記号は，“x，b，o” が 0 に，“y，a” が 1 に，“k” が 2 のようにインデキシング記号に置き換えられる．先の “kxx” は “200” に変換される．最後に，インデキシング記号に変換された記号列中に，決定木を構成するノード “20” が存在すればスパムメール，存在しなければ正規メールに分類されることになる．

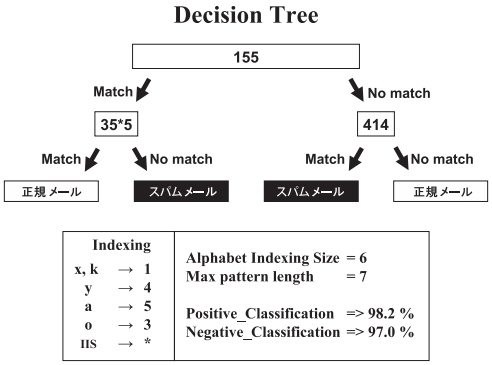


図 3 BONSAI が出力した電子メール判定の決定木
Fig. 3 A decision tree obtaind from BONSAI to classify emails.

3.5 BONSAI によるインデキシングおよび決定木

図 3 は，学習例を構成するアルファベット {x，y，a，b，k，o} の 6 つの記号に対して，BONSAI がインデキシングを行い，学習例であるスパムメール群と正規メール群を最も効率良く分類できる規則を決定木として出力した結果である．BONSAI の学習パラメータとして，Indexing Size を 6，Max Pattern Length を 7 に設定した．Indexing Size は，学習例に含まれる記号に対して，インデキシングに使う文字の最大文字数，Max Pattern Length は，決定木を構成しているノードのパターンの最大長を規定している．BONSAI は x，y，a，k，o の各文字に対し，インデキシングによって x と k を 1，y を 4，a を 5，o を 3 と分類した．残りの b は*で表現された部分のみで許容される．*は Insignificant indexing symbol であり [16]，学習例を構成する x，y，a，b，k，o すべての記号がノードパターン中の*部分で許容されることを意味する．つまり，Insignificant indexing symbol で表現された部分のみで許容される記号は，BONSAI がインデキシング処理を行う際に，ノードパターンを構成する重要な位置には関与しない，あるいは特徴配列には関与しない記号であることを意味している．図 3 の決定木のノードパターン 35*5 を例にとると，o に始まり，続いて a，次にすべての記号 x，y，a，b，k，o のうちの 1 つ，最後に a で終わる記号列を示しており，正規メール群によく出現する文字列で構成される特徴パターンを表している．メールの分類を行う場合には，分類対象メールも学習例と同様に単語の出現頻度による記号列変換を行った後，BONSAI の出力したインデキシング記号に置き換えられ，決定木に含まれるパターンが記号列変換後のメール内に存在するか，しないかで判定を行う．図 3 の場合，記号列変換後のメール内に 155 および 35*5 が含まれる場合と 155 および 414 が含まれない場合，正規メールと判定される．155 は含まれるが 35*5 は含まれない場合と 155 は含まれないが，414 は含まれる場合は，スパムメールと判定される．

4. ベイジアンフィルタ

ベイジアンフィルタはスパムメール群と正規メール群を学習例として，それぞれのメール群に出現した各単語とその出現頻度から，それぞれの単語がスパムメールに出現する確率（スパム確率）を算出する．判定対象とするメールに出現した各単語に対する単語のスパム確率をもとに，そのメールがスパムメールである確率を算出する．メールのスパム確率があらかじめ定めた閾値以上であればスパムメール，そうでなければ正規メールと判定する [17]．ベイジアンフィルタの 1 つである bsfilter は，メール判定の閾値を 0.582 以上でスパムメール，そうでなければ正規メールと判定するように設定されている [6]．bsfilter は正規メールの正解率は高く，スパムメールの正解率が低いという特徴があるが，これは閾値の設定が大きく影響している．

5. BONSAI と bsfilter による併用フィルタリング

BONSAI と bsfilter の両方を用いて，スパムメールと正規メールを分類する併用フィルタリングについて説明する．2 つのフィルタはそれぞれが異なる判定方式を採用しているため，一方がメールを誤判定しても，もう片方が正しく判定できる可能性がある．BONSAI を用いたフィルタリング手法は，3.3 節で説明したように，学習群のメール中に含まれる単語の出現頻度と語順により分類規則を作成し，メール中に特定の記号列のパターンが含まれているかにより判定を行う．つまり，BONSAI は図 1 b で示したように，決定木のノードパターンにマッチするかによってのみ判定を行うため，メールの特徴を強く表すパターンの記号列が判定に利用される [7]．一方，bsfilter は学習群のメール中に含まれる各単語の出現頻度によって，総合的にそのメールがどれぐらいスパムメールであるかの確率を算出している [17]．いい換えれば，bsfilter はメール全体の文字列情報から，メールの全体的な類似傾向を利用して判定を行っている．すなわち，BONSAI はメールの局所的な情報を利用しているのに対して，bsfilter はメールの平均的な情報を利用しているため，判定方式が異なっていることが提案する併用フィルタリングに非常に重要な点である．また，語順を判定に利用する BONSAI を併用することで，ベイジアンフィルタの回避策として使用されているワードサラダ [18] にも対応できる可能性がある．

先行研究として統計的手法と事例ベース手法を併用したスパムフィルタリング [10] 手法がある．新たな傾向のメールに素早く対応できる事例ベースフィルタと長期的な傾向に基づいた高精度な分類を実現する統計的フィルタの特徴をあわせ持つシステムとして，高い分類精度を実現している．一方，本論文で提案する手法は，bsfilter と BONSAI を用いた分類を連続して行う．BONSAI は決定木によっ

表 3 閾値変更前の bsfilter の判定結果
Table 3 Accuracy rate evalution of bsfilter before changing the evaluation value.

$trainingmail$	1,000					
$System$	bsfilter[0.582]	BONSAI	併用 I	併用 II	併用 III	併用 IV
$test_{total}$	2,000					
$test_{spam}$	1,000					
$test_{ham}$	1,000					
$correct_{spam}$	746	969	745	970	745	970
$error_{spam}$	254	31	255	30	255	30
$correct_{ham}$	1,000	979	1,000	979	1,000	979
$error_{ham}$	0	21	0	21	0	21
$AccuracyRate_{spam}$	74.6	96.9	74.5	97.0	74.5	97.0
$AccuracyRate_{ham}$	100	97.9	100	97.9	100	97.9

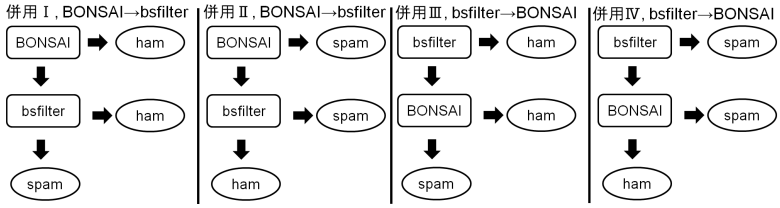


図 4 BONSAI と bsfilter による併用フィルタリングの流れ
Fig. 4 Process of Spam Mail Filtering System with BONSAI and bsfilter.

てメールを分類するため，新たな傾向のメール群やその他の複数の傾向を持つメール群を 1 つの決定木で同時に判別できる．しかも，決定木の出力が完了すれば，事例ベースフィルタのような学習メールの増大による処理時間の増大は起こらない．また，bsfilter と BONSAI を用いた分類を連続して行うことで，一方が誤分類したメールをもう一方で回収し，誤分類を低減させることができる可能性が高い．これらの特徴は，これまでにない新たなメールフィルタリングシステムとして，高い分類精度を実現できると考えられる．

bsfilter と BONSAI を用いたそれぞれのフィルタリング性能については，後述の**表 3** を見ると分かるように，BONSAI は正規メールおよびスパムメールいずれの正解率も高いが，少数の正規メールを誤判定する傾向にあり，一方，bsfilter は正規メールにスパムメールを多く混入してしまう傾向にあるが，正規メールの正解率は，100%に近い．そこで，併用フィルタリングでは，2 つの手法の長所を組み合わせ，弱点を補い合って分類精度の高いフィルタリング手法を確立することを目的とする．**図 4** に bsfilter と BONSAI の考えられる併用方法を示す．ham は正規メール，spam はスパムメールを表す．

図 4 の併用 I は，1 段階目の BONSAI が正規メールと判定したメールに対してはその時点で判定を決定し，スパムメールと判定したメールに対しては 2 段階目の bsfilter が判定を行う．1 段階目の BONSAI により正規メールと誤判定されるスパムメールは少ないが，2 段階目の bsfilter により，正規メールと誤判定するスパムメールが増えてしまう．

併用 II は 1 段階目の BONSAI がスパムメールと判定したメールに対してはその時点で判定を確定し，正規メールと判定したメールに対しては 2 段階目の bsfilter が判定を行う．1 段階目の BONSAI がスパムメールと判定したメール群に含まれた正規メールは回収できないことが大きな欠点である．併用 III は 1 段階目の bsfilter が正規メールと判定したメールに対してはその時点で判定を確定し，スパムメールと判定したメールに対しては 2 段階目の BONSAI が判定を行う．bsfilter は 100%の正規メールを回収することが期待できるが，正規メールと判定したメールの中に大量のスパムメールが含まれてしまうことが難点である．併用 IV は 1 段階目の bsfilter がスパムメールと判定したメールに対してはその時点で判定を確定し，正規メールと判定したメールに対しては 2 段階目の BONSAI で判定を行う．bsfilter はスパムメールの分類は正しく行うが，分類しきれないスパムメールを正規メール群に混入させてしまうため，正規メールの判定が BONSAI によることになり，少数のスパムメールと少数の正規メールを誤判定してしまう危険性が残る．

6. 分類判定結果による bsfilter の閾値変更

本章では bsfilter，BONSAI によるメールフィルタ，およびこれらの併用フィルタで判定を行った結果を示す．bsfilter の分類を決めるパラメータである閾値は，まずは通常使われる値に設定する．この結果をもとに，併用フィルタによる問題点を検討し，メールのスパム確率の分布により bsfilter の閾値を変更する方法を提案する．

6.1 使用メールサンプルデータ

実験には，2007 TREC Public Spam Corpus（TREC07）[19] の partial データセットを使用した．以前の研究 [7] では，日本語の定型文を多く含むメールデータを用いていたが，一般性の低いメール群であったため，提案手法の効果の確認にはこれまでのスパムメールフィルタ研究の検証実験で実績のある [10], [11], [12], [13] TREC07 の英文メールデータを用いている．TREC07 の partial データセットは 2007 年 4 月 8 日から 2007 年 7 月 6 日のメール 30,338 通（スパムメール 6,280 通，正規メール 24,058 通）から構成される．メールは時系列順に収められており，実験を行う際も時系列順に取り扱った．partial データセットの時系列の先頭からスパムメール，正規メールそれぞれ 1,000 通ずつ選出し，学習用メール群とした．学習数に応じて，学習用メール群からメールを選出した．さらに，時系列順にスパムメール 1,000 通，正規メール 1,000 通を分類対象メール群として判定を行った．なお，メールの添付ファイルは除外したが，メールヘッダの情報は分類に用いた．

6.2 bsfilter の閾値変更前の判定結果

学習数を 1,000 通（スパムメール 500 通，正規メール 500 通）として判定を行った結果を表 3 に示す．学習メール数は表 3 中の「*training mail*」で表す．分類対象メール群を分類したシステム名を「*System*」で示す．bsfilter の後に書かれた数値 0.582 はメール判定に用いられる閾値の値である．BONSAI によるメールフィルタリングは「BONSAI」，bsfilter と BONSAI の併用フィルタリングは「併用」，また，「併用」の後に書かれている番号は図 4 の併用フィルタの番号と対応している．分類対象メール数，分類対象スパムメール数，分類対象正規メール数はそれぞれ「$test_{total}$」，「$test_{spam}$」，「$test_{ham}$」で表す．$correct_{spam}$($correct_{ham}$) は「スパムメール（正規メール）を正しく分類したメール数」である．$error_{spam}$($error_{ham}$) は「スパムメールを正規メール（正規メールをスパムメール）と誤って分類したメール数」である．$AccuracyRate_{spam}$ は「スパムメールを正しく分類した割合」であり式 (7) から，$AccuracyRate_{ham}$ は「正規メールを正しく分類した割合」であり，式 (8) から算出する．

6.3 併用の組合せによる結果の考察

理想のスパムメールフィルタは，すべてのスパムメールと正規メールを間違いなく分類する精度を持つものであるが，現実的に最も重要とされる精度は，正規メールの正解率である．つまり，多少のスパムメールが混入したとしても，まずはすべての正規メールを間違いなく回収する能力が最重要であり，その次に，いかにしてスパムメールの混入を防ぐかを考えなければならない．

併用 I の場合，1 段階目の BONSAI で判定を行った際，少数の正規メールをスパムメールと誤判定してしまったが，2 段階目の bsfilter は正規メールの正解率が高いのでこれらの誤判定した正規メールを正しく回収した．しかし，スパムメールの正解率が低いので，BONSAI が正しく判定したスパムメールのうちの多くを誤判定してしまった．すなわち，2 段階目で分類されるメールのほとんどはスパムメールであるため，スパムメールの判定は bsfilter に依存してしまい，低いスパムメールの正解率となってしまっている．

併用 II の場合，1 段階目の BONSAI での分類では，BONSAI がスパムメールと判定したメールの中に少数の正規メールが混入してしまった．1 段階目の BONSAI がスパムメールと判定したメールに対しては，この時点で判定が確定するので，ここで誤判定された正規メールは回収することができない．2 段階目の bsfilter で分類されるメールは，ほとんどが正規メールである．bsfilter は正規メールの正解率が高いので BONSAI が正規メールと判定したメールをそのまま正しくを回収できている．

併用 III の場合，1 段階目の bsfilter での分類では，bsfilter は正規メールをすべて正しく判定したが，正規メールと判定したメールの中に大量のスパムメールが含まれてしまっている．2 段階目の BONSAI は，正規メールの回収と混入したスパムメールを取り除くための判定を行うことになる．今回の実験では，1 段階目で正規メールがすべて回収できているため，ここでの BONSAI の処理はすべてのメールをスパムメールに分類しなければならない．しかし，一部を正規メールと誤判定してしまったために，わずかではあるが最終的なスパムメールの正解率は，bsfilter のみの場合に比べて減少した（74.6%から 74.5%）．

併用 IV の場合，1 段階目の bsfilter での分類では，bsfilter がスパムメールと判定したメールの中に正規メールは含まれなかった．しかし，bsfilter がスパムメールと判定したメール数は少ない．また，すべての正規メールが 2 段階目へ送られるため，正規メールの判定は BONSAI に依存してしまい，少数のスパムメールおよび正規メールを誤判定してしまう BONSAI では，すべての正規メールを回収することはできなかった．以上 4 つの比較により，正規メールの正解率を可能な限り 100%に近づけるためには，正規メールに対しての判定を 2 回行う併用 I と III が有効であると考えられる．さらに，この 2 つの併用方法において，スパムメールの正解率を高くするためには bsfilter のスパムメールの誤判定を減少させる必要がある．

6.4 メールのスパム確率の分布

bsfilter のスパムメールの誤判定を減らすために，メール判定の閾値を通常使われる値より低く設定する．bsfilter のメール判定に使用される閾値は 0.582 以上でスパムメー

ルに分類されるように初期設定されている [6]．これは，正規メールの正解率を高くするため閾値を高く設定していると考えられ，その結果，多くのスパムメールを正規メールと判定している．つまり，閾値を下げることにより，スパムメールの正解率は上昇するが，反対に正規メールの正解率は減少してしまう．そこで，メールの閾値を変更するために各メールのスパム確率がどのように分布しているかを調べた．各メールのスパム確率の分布を**表 4** に示す．変数 a がこの閾値に用いられる値であり，この値によってスパムメール分類の確かさが決められている（メールのスパム確率）．スパムメール 1,000 通，正規メール 1,000 通の分類対象メール群について，スパム確率を 1 から 0 の間で 0.1 ずつ 10 個に区切り，各メールがどの範囲に含まれているかを示した．ただし閾値の初期値である 0.582 についてはより詳しく分類の境界を示している．

6.5 メール判定の閾値の変更

表 4 より，スパムメールの分布ではメールのスパム確率が 0.9 以上の範囲に含まれるメール数が 698 通と最も多い．正規メールの分布では，0.1 未満の範囲に含まれるメール数が 963 通と最も多く，正規メール全体の 96.3%を占めている．メール判定の閾値が初期値の 0.582 では，多くのスパムメール（254 通）が正規メールと判定されてしまい，スパムメールの誤判定を引き起こしている．そこで，スパムメールの誤判定を減少させるために閾値の値を従来よりも低く設定する．正規メールのほとんどがメールのスパム確率が 0.1 未満であり，スパムメールの誤判定をできるだけ少なくする理由から，今回は閾値を 0.1 以上でスパムメールと判定することにした．最適な閾値の選択法については，今後の課題とする．

6.6 bsfilter の閾値変更後の結果と併用 III の適用

表 4 より，bsfilter の閾値を 0.1 以上でスパムメールと設定した場合，a が 0.1 未満である，スパムメール 12 通，正規メール 963 通が正規メールと判定され，a が 0.1 以上である，スパムメール 988 通，正規メール 37 通がスパムメールと判定される．正解率はスパムメールが 98.8%，正規メールが 96.3%となる．閾値変更前と比較して，スパムメールの正解率が 24.2%上昇し，スパムメールの誤判定を大幅に減少させることができる．しかし，正規メールの正解率は閾値変更前と比較して 3.7%減少してしまう．併用 I および併用 III では，1 段階目で正規メールと判定されたスパムメールについては 2 度と回収できないため，BONSAI よりもスパムメールの誤判定が少ない bsfilter を 1 段階目に使用する併用 III を併用フィルタとして用いる．先に述べた，1 段階目の bsfilter の分類では，多くの正規メールを回収し，その中に含まれるスパムメールの数を減少させることができる．一方，BONSAI の正規メール，スパムメールともに正解率が高いという特徴を利用して [20]，bsfilter が誤判定した正規メールを 2 段階目で回収することが期待できる．

7. 併用フィルタリングの結果と考察

閾値変更後の併用フィルタリングの分類結果を**表 5** に示す．表 5 には，bsfilter の閾値を 0.1 以上でスパムメールと判定した結果，BONSAI を応用したメールフィルタリングで判定した結果，bsfilter と BONSAI を併用した結果を示す．

bsfilter のみの場合，スパム確率が 0.582 以上でスパムメールと判定する基準から，0.1 以上でスパムメールと判定する基準に変更した結果，正規メールの正解率は 96.3%となり，閾値変更前の 100%と比較して正解率が減少した．しかし，BONSAI と組み合わせることで，bsfilter が誤判定した正規メール 37 通中 36 通を正しく判定しており，99.9%という正規メールの正解率を得た．最終的に誤分類してしまった 1 通の正規メールは，本文に文章をほとんど含んでおらず，文章自体も正規メールとスパムメールのどちらにもよく出現する簡単な単語で構成されていたため，

表 4 メールのスパム確率の分布
Table 4 Distribution of spam mail rate.

範囲	spam	ham
$0.9 \leq a \leq 1.0$	698	0
$0.8 \leq a < 0.9$	19	0
$0.7 \leq a < 0.8$	7	0
$0.6 \leq a < 0.7$	16	0
$0.582 \leq a < 0.6$	6	0
$0.5 \leq a < 0.582$	223	2
$0.4 \leq a < 0.5$	14	9
$0.3 \leq a < 0.4$	2	6
$0.2 \leq a < 0.3$	1	9
$0.1 \leq a < 0.2$	2	11
$0.0 \leq a < 0.1$	12	963

表 5 閾値変更後の bsfilter の判定結果
Table 5 Result of an accuracy of bsfilter after changing the evaluation value.

$trainingmail$	1,000		
$System$	bsfilter[0.1]	BONSAI	併用フィルタ
$test_{total}$	2,000		
$test_{spam}$	1,000		
$test_{ham}$	1,000		
$correct_{spam}$	988	969	961
$error_{spam}$	12	31	39
$correct_{ham}$	963	979	999
$error_{ham}$	37	21	1
$accuracy\ rate_{spam}$	98.8	96.9	96.1
$accuracy\ rate_{ham}$	96.3	97.9	99.9

正規メールとして判断するために十分な情報が得られなかったと推測される．我々が提案するスパムメールフィルタは，併用するフィルタのそれぞれが異なる基準で判定を行うことにより，一方が誤判定してしまった正規メールをもう一方が正しく判定できる可能性を示している．1段階目の bsfilter ではスパムメールの誤判定を可能な限り少なくすることが必要であるが，これを bsfilter の閾値を従来よりも低く設定することにより実現した．BONSAI を応用したメールフィルタを2段階目に配置することで，判定方式の違いから bsfilter が誤判定した正規メールを回収することができた．つまり，スパムメールの正規メールへの誤分類を最小限に抑えつつ，正規メールの正解率を上昇させることが可能となった．

電子メール利用者にとって最も重要なのは正規メールの正解率であるが，スパムメールの数が増加している現在では，スパムメールの正解率も重要となってきている．本論文で提案した BONSAI と bsfilter の併用フィルタリングは電子メール利用者のニーズに応えるメールフィルタであるといえる．

8. おわりに

本論文では，BONSAI と bsfilter を組み合わせ，さらに bsfilter のメール判定に用いられる閾値を従来よりも低く設定することで正規メールとスパムメールともに，正解率の高いメールフィルタリング手法を検討した．併用フィルタの組合せ方の決定については，実際に4つの場合の結果を比較し，それぞれの組合せについて実験データを基に考察し，利用者にとって最も重要な，正規メールを最もよく回収できる組み合わせ方式を採用した．bsfilter の閾値の変更については，スパムメールの誤判定をできる限り少なく抑えるために，閾値を従来より低く設定することで BONSAI との併用効果を向上させることができた．

今後は，わずかであるが低下した正規メールの正解率（100%→ 99.9%）を回復させるため，bsfilter の閾値の最適化手法などの開発に取り組むと同時に，誤判定したメールの追加学習の効果などの検討を行い，さらに高性能なメールフィルタの実現を目指していきたい．

謝辞 本研究の一部は，科学研究費補助金若手研究（B）（21700078）の援助を受けて実施した．

参考文献

[1] 安東孝二：世界の電子メールを spam 制御へ，情報処理，Vol.46, No.7, pp.741–746 (2005).
[2] 総務省：迷惑メールへの対応の在り方に関する研究会最終とりまとめ (2008)，入手先 ⟨http://www.soumu.go.jp/main_sosiki/joho_tsusin/policyreports/chousa/mail_ken/index.html⟩.
[3] Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E.: A Bayesian Approach to Filtering Junk E-Mail, Learning for Text Categorization: Papers from the 1998 Workshop, AAAI Technical Report WS-98-05 (1998).
[4] Graham, P.: A Plan for Spam (2002), available from ⟨http://www.paulgraham.com/spam.html⟩.
[5] Graham, P.: Better Bayesian Filtering (2003), available from ⟨http://www.paulgraham.com/better.html⟩.
[6] bsfilter-bayesian spam filter, available from ⟨http://sourceforge.jp/projects/bsfilter/⟩.
[7] 杉井 学，松野浩嗣：機械学習によるスパムメールの特徴の決定木表現，情報処理学会研究報告 2007-DPS-130(16), pp.183–188 (2007).
[8] Shimozono, S., Shinohara, A., Shinohara, T., Miyano, S., Kuhara, S., and Arikawa, S.: Knowledge Acquisition from Amino Acid Sequences by Machine Learning System BONSAI, *Trans. Inform. Process. Soc. Japan*, Vol.35, pp.2009–2018 (1994).
[9] Usuzaka, S., Kim, L.S., Tanaka, M., Matsuno, H. and Miyano, S.: A Machine Learning Approach to Reducing the Work of Experts in Article Selection from Database: A Case Study for Regulatory Relations of S. cerevisiae Genes in MEDLINE, *Genome Informatics*, Vol.9, pp.91–101 (1998).
[10] 鳴海建太，西田京介，山内康一郎：統計的手法と事例ベース手法を併用したスパムメールフィルタリング，電子情報通信学会論文誌 D，Vol.J91-D, No.11, pp.2569–2578 (2008).
[11] 北村祐貴，狩野 均：事前処理に k-means 法を利用したスパムフィルタの開発，情報処理学会研究報告 2009-MPS-76, No.12, pp.1–8 (2009).
[12] Byun, B., Lee, C., Webb, S., Irani, D. and Pu, C.: An Anti-spam Filter Combination Framework for Text-and-Image Emails through Incremental Learning, *Proc. 4th Conference on Email and Anti-Spam* (*CEAS 2009*), Mountain View, CA (July 2009).
[13] Jou, C. and Shih, Y.: A Spam Email Classification Based on the Incremental Forgetting Bayesian Algorithm, *Business and Information 2012*, D826-838 (2012).
[14] KAKASI-漢字→かな（ローマ字変換）プログラム，入手先 ⟨http://kakasi.namazu.org/⟩
[15] Robinson, G.: A statistical approach to the spam problem, Linux Journal archive, Vol.2003, No.107 (Mar. 2003), available from ⟨http://www.linuxjournal.com/article/6467⟩.
[16] Sugii, M., Okada, R., Matsuno, H. and Miyano, S.: Performance Improvement in Protein *N*-Myristoyl Classification by BONSAI with Insignificant Indexing Symbol, *Genome Informatics*, Vol.32, pp.277–286 (2007).
[17] 小川健司，稲葉宏幸：記号と未知語の分布を用いたベイジアンスパムフィルタの提案，電子情報通信学会技術研究報告，2009-IA-108(460), pp.209–212 (2009).
[18] 田端利宏：SPAM メールフィルタリング：ベイジアンフィルタの解説，情報の科学と技術，Vol.56, No.10, pp.464–468 (2006).
[19] TREC 2007 Public Corpus, available from ⟨http://plg.uwaterloo.ca/g̃vcormac/treccorpus07/⟩.
[20] 角 朝香，日下野隆謙，杉井 学，松野浩嗣：機械学習を応用したスパムメールフィルタリング手法の検討と評価，マルチメディア通信と分散処理ワークショップ論文集，pp.201–204 (2009).

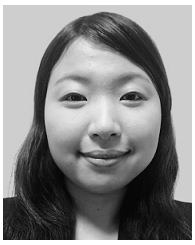
推薦文

推薦論文は，実用的なスパムメールフィルタリング手法の提案をしている．この手法は，ベイジアンフィルタによってスパムメールであると分類されたメールに対して，著者らが独自に開発を進めてきた機械学習システムを用いたメールフィルタを組み合わせて用いる．これは「メール中の単語の出現頻度」による分類と，「メール中の重要な単語とその単語の出現順」による分類という，異なる判定の効果的な組み合わせ方を提示している．さらに，ベイジアンフィルタでのメール判定に用いられる閾値の変更を行うことで，正規メールとスパムメールともに分類正解率が高いメールフィルタリング手法を検討している．情報処理学会論文誌に推薦するにふさわしい論文と判断される．

（情報処理学会中国支部長　三池秀敏）

山口 博之

1986 年生．2010 年山口大学理学部物理情報科学科卒業．2012 年同大学大学院理工学研究科修士課程修了．同年株式会社日立システムズ入社．



角 朝香

1985 年生．2008 年山口大学理学部物理情報科学科卒業．2010 年同大学大学院理工学研究科修士課程修了．同年株式会社日本ラッド入社．

杉井 学　（正会員）

1972 年生．1996 年山口大学理学部生物学科卒業．1998 年同大学大学院理学研究科修士課程修了．2001 年同大学院理工学研究科博士後期課程修了．博士（理学）．2001 年同大学ベンチャービジネスラボラトリー非常勤研究員．2002 年同大学大学情報機構メディア基盤センター助手．2006 年同大学准教授．機械学習を用いたゲノム情報解析やネットワークシステム開発に従事．

松野 浩嗣　（正会員）

1960 年生．1982 年山口大学工学部電子工学科卒業．1984 年同大学大学院修士課程修了．1984～1987 年山口短期大学，1987～1994 年大島商船高等専門学校勤務．1997 年山口大学理学部助教授．2005 年同大学教授．2006 年同大学大学院理工学研究科教授．計算機ネットワーク構築技術と生命のシステム的理解に関する研究に従事．理学博士．IEEE，電子情報通信学会各会員．