情報処理学会論文誌　数理モデル化と応用　Vol. 1　No. 1　175–184 (Sep. 2008)

特徴抽出方法の改善による ベイジアンフィルタの精度向上

谷岡広樹[†1]　中川尚[†1]　丸山稔[†2]

本稿では，従来法の 1 つであるベイジアンフィルタを用いた spam メールフィルタの精度（true negative rate）を改善する方法について提案する．これまでの学習型 spam メールフィルタとしては，ベイジアンフィルタがよく利用されており，一定の成果が得られている．しかしながら，ベイジアンフィルタを利用した方法においても，誤検出率（false positive rate）の低減や，さらなる精度向上が期待される．我々は，単語の spam 確率（尤度）の分布およびメールの spam 度の分布状況を分析し，誤検出をおさえながらも，高い判定精度を実現する方法について提案し，その精度について，従来方式と比較して評価する．

Improvement of Feature Extraction for Bayesian Spam Filtering

Hiroki Tanioka,[†1] Takashi Nakagawa[†1] and Minoru Maruyama[†2]

We propose an improved baysian filter for spam mail detection. Bayesian filter was used on existing learning spam filters which achieved some positive results. Although we expect them to improve the true negative rate while keeping the false positive rate low. Therefore, it was based on a thorough review of distribution for each word and mail that our means of spam mail detection showed an impressively higher accuracy than ever.

†1 株式会社ジャストシステム
JustSystems Corporation
†2 信州大学工学部情報工学科
Department of Information Engineering, Faculty of Engineering, Shinshu University

1. はじめに

近年，日々刻々と変化する spam メールに対抗し，より高精度な spam メールフィルタを実現するために，さまざまな手法が研究[9] されているが，我々は，新手の spam メールに対しては，ユーザからのフィードバックにより追加学習することで，ユーザ環境に応じて，判定精度を高精度に保つことができると考える．また，日常的に利用する電子メールの誤検出（false positive）は，見逃し（false negative）よりも，ユーザに与える影響が大きいと考え，我々の目指す spam メールフィルタの要件は，以下の 2 つとする．

(1)　追加学習が高速であること．
(2)　誤検出が十分に少ないこと．

すでに，文書フィルタリングに応用可能な分類器には，さまざまなアルゴリズムや学習モデルが提案されているが，我々は，**追加学習が高速である**ために，学習コストの低いアルゴリズムとして，ベイジアンフィルタを学習モデルに採用する．また，ベイジアンフィルタによる**誤検出を十分に少なくする**ため，閾値（threshold）を適切に設定する．このときの閾値の設定基準は，誤検出が十分に少ない条件として，誤検出率（FPR; false positive rate）が一定の値以下の場合とする．本稿では，特徴抽出の方法を改善することで，以上の条件を満たしながら，従来手法と比較して精度（TNR; True Negative Rate）の高い spam メールフィルタを実現する．

1.1 spam メールフィルタ

従来方式には，サポートベクターマシン（SVMs; Support Vector Machines）を用いる方法や，ロジスティック回帰（Logistic regression）を用いて，より厳密に統計量を計算する方法等が提案されているが，本稿では，**追加学習が高速であること**を重視するため，比較的計算量の小さいアルゴリズムとして，ベイジアンフィルタを採用する．

spam メールフィルタのためのベイジアンフィルタは，Graham の文献[2] をきっかけに，広く普及しているが，PaulGraham 方式やその改良版の Robinson[4] 方式においても，誤検出率は低いままに，より高い spam 判定精度を実現することが望まれる．そのため，本稿ではまず，ベイジアンフィルタを用いた従来方式について概観し，単語の spam 確率やメールの spam 度の分布状況を調べる．その後，従来方式の課題を整理し，その課題を改善する方法を提案する．

1.2　データセットと実験環境

本稿では，すべての実験の spam メールコーパスとして，以下のデータセット[⋆1]を用いる．

- spam メール：1,671 通
- ham [⋆2]メール：3,260 通

文字コードによる内訳は，spam メール（日本語 209 通，英語 1,119 通，その他 343 通），ham メール（日本語 2,316 通，英語 595 通，その他 349 通）である．なお，本文，ヘッダ情報，添付ファイルを含むすべてのメールデータを形態素解析し，その結果得られた単語の表記を，特徴データ（feature）とするよう前処理を行った．また，すべての spam メールフィルタは，Java 言語（JDK1.5.0_11-b03）を用いて実装した．

2.　ベイジアンフィルタ

ベイジアンフィルタを用いた現行の spam メールフィルタ[10] の基礎技術について概観すると，以下のようなものがある．

- 実用的 spam メールフィルタとして最もよく知られている典型的手法が Paul Graham により提案された方法（PaulGraham）．
- PaulGraham 方式の（技術的，理論的）基礎を与えるのが naive Bayes classifier（NaiveBayes）．
- PaulGraham 方式の公表以後，種々の改良が加えられており，その典型例が Gary Robinson により提案された方法（Robinson，Robinson-Fisher）．

NaiveBayes 方式は，各単語が互いに独立であることを仮定し，各単語の spam 確率の同時確率によってメールの spam 確率を定義するが，実際には，各単語は互いに独立ではないため，PaulGraham 方式では，特徴的な単語を n 語抽出するヒューリスティックを用い，Robinson 方式では，訓練データの統計量から，メールの spam 度を計算する．本章では，現行手法の基礎と，その問題点を明らかにするために分析を行う．

2.1　NaiveBayes 方式

NaiveBayes 方式（naive Bayesian classifier）[1),8)] では，N 語の単語を含むメールの各単語の特徴量 w_i が，クラス c が決まっているという条件で独立であるとき，次式を最大化するカテゴリ $\hat{c}$ を選べばよい．

$$\hat{c} = \arg\max_{c} P(c) \prod_{i=1}^{N} P(w_i|c) \tag{1}$$

具体的には，ある単語 w_i が spam として登場した回数を s_i，ham として登場した回数を h_i，spam メールの総数を S，ham メールの総数を H としたとき，ある単語 w_i の spam 確率 $p(w_i)$ を次式で表す．

$$p(w_i) = \frac{s_i/S}{h_i/H + s_i/S} \tag{2}$$

また，単語 $w_1, \cdots, w_N$ を含むメール M の spam 確率 $P(S)$ および ham 確率 $P(H)$ について，それぞれ演算誤差を防ぐために対数を用いて計算する．

$$P(S) = \sum_{i=1}^{N} \log(p(w_i)),$$

$$P(H) = \sum_{i=1}^{N} \log(1 - p(w_i))$$

さらに，メール M の spam 度 $P(M)$ を

$$P(M) = \frac{P(H)}{P(S) + P(H)} \tag{3}$$

とすると，spam 度 $P(M)$ が閾値 t を上回った場合，メール M は spam であると判定できる．

このとき閾値 t は，事前に分布状況を調べておいて，精度と誤検出率のバランスを考えて任意の値を設定する必要がある．実際には，2.2 節で分析した結果を用いる．

2.2　NaiveBayes 方式の分析

図 1 は，NaiveBayes 方式を用いた場合の，spam 確率に対する平均単語数の分布を表している．横軸には，NaiveBayes 方式に基づいて計算した spam 確率 $p(w_i)$ を，縦軸には，spam 確率を 0.05 ずつの区間に分けて，各区間に含まれる平均単語数（単語数 ÷ メール数）をとったヒストグラムである．ここで，spam 確率が $(0.9, 1]$ の範囲には，spam メールに含まれる単語が多く分布し，spam 確率が $[0, 0.1]$ の範囲には，ham メールに含まれる単語が多く分布するが，それ以外の範囲 $[0.1, 0.9]$ にも分布していることが確認できる．

図 2 は，NaiveBayes 方式を用いた場合の，spam 度に対するメール数の分布を表している．横軸には，spam 度 $P(M)$ を，縦軸には，spam 度を 0.05 ずつの区間に分けて，各区

⋆1 2007 年 8 月の約 2 週間の間に，個人のメールアドレスで受信したメールである．本データセットの入手方法に関する問合せ先：hiroki.tanioka@justsystems.com

⋆2 ham は非 spam（non spam）を意味する．

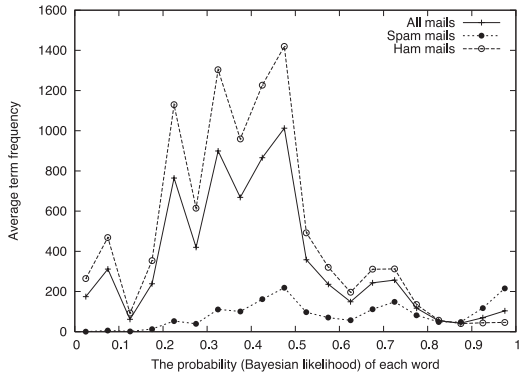

図 1　spam 確率に対する平均単語数の分布 – NaiveBayes
Fig. 1　The distribution of average term frequencies for each probability on the Naive Bayes method.

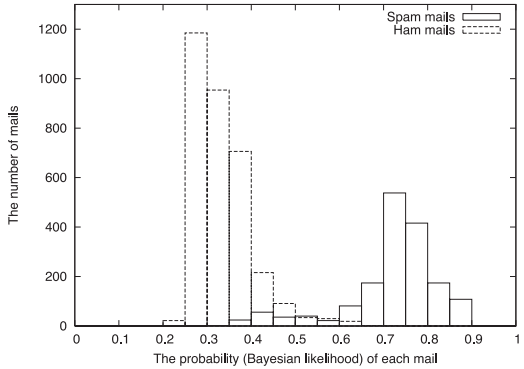

図 2　spam 度に対するメール数の分布 – NaiveBayes
Fig. 2　The distribution of mails for each probability on the Naive Bayes method.

間に含まれるメール数をとったヒストグラムである．ここで，ham メールについてプロットしたグラフ（Ham mails）は $[0, 0.5]$ の範囲に多く分布し，spam メールについてプロットしたグラフ（Spam mails）は $(0.6, 1]$ の範囲に，多く分布している．この分布状況から，NaiveBayes 方式での閾値 t の範囲を $[0.5, 0.7]$ とする．

2.3　PaulGraham 方式

PaulGraham 方式は，Paul Graham によって提案された方式[2),3)] である．ある単語 w_i が spam として登場した回数を s_i，ham として登場した回数を h_i，spam メールの総数を S，ham メールの総数を H としたとき，ある単語 w_i の spam 確率 $p(w_i)$ は，

$$p(w_i) = \frac{s_i/S}{a \cdot h_i/H + s_i/S} \tag{4}$$

である．ここで，a は誤検出を低減することを狙ったバイアスである．

また，単語 $w_1, \cdots, w_N$ を含むメール M に含まれる単語の中から，spam 確率が 0.5 から離れている順に n 語を選び（$|p(w_i) - 0.5|$ の大きい順にソートした後，上位 n 語を抽出し），メール M の spam 確率 $P(S)$ および ham 確率 $P(H)$ を，

$$P(S) = \prod_{i=1}^{n} p(w_i),$$

$$P(H) = \prod_{i=1}^{n} (1 - p(w_i))$$

とする．さらにメール M の spam 度 $P(M)$ を

$$P(M) = \frac{P(S)}{P(S) + P(H)} \tag{5}$$

とすると，spam 度 $P(M)$ が閾値 t を上回った場合に，メール M は spam であると判定できる．このとき，文献 2) に準拠すると，$a = 2$，$n = 15$，$t = 0.9$，単語が未知語の場合は $p(w_i) = 0.4$ である．

2.4　PaulGraham 方式の分析

図 3 は，PaulGraham 方式で閾値 t は 0.9 とし，10-folds の交差検定により計測した場合の，単語数 n と精度（TNR）の関係を表す．文献 2) では，単語数 n は 15 語とするとあったが，実験の結果，単語数 n が 7 のとき 0.98337 と，最も高い精度のため，$n = 7$ とする．

図 4 は，PaulGraham 方式を用いた場合の，spam 確率に対する選択された平均単語数の分布を表している．横軸には，spam 確率 $p(w_i)$ を，縦軸には，spam 確率を 0.05 ずつの区間に分けて，選択された単語が，各区間に含まれる平均単語数（単語数 ÷ メール数）をとったヒストグラムである．ここで，spam 確率が $(0.9, 1]$ の範囲には，spam メールに含まれる単語が多く，spam 確率が $[0, 0.1]$ の範囲には，ham メールに含まれる単語が多く分布していることが確認できるが，spam 確率が $[0.1, 0.9]$ の範囲には，ほとんど単語の分布が

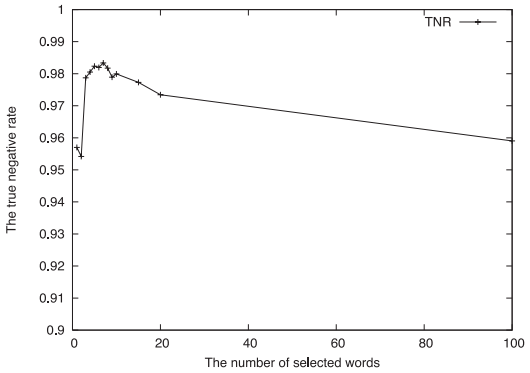

図 3　選択する単語数（n）に対する精度（TNR）— PaulGraham
Fig. 3　The true negative rate for the number of selected words on the PaulGraham method.

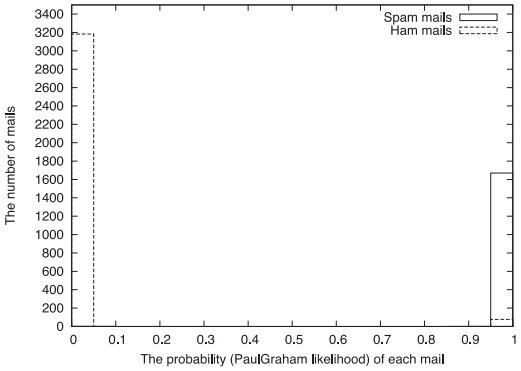

図 5　spam 度に対するメール数の分布 – PaulGraham
Fig. 5　The distribution of mails for each probability on the Paul Graham method.

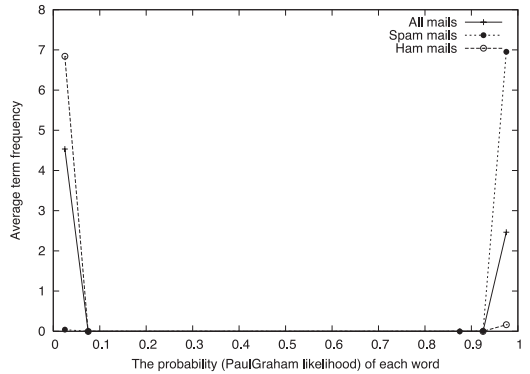

図 4　spam 確率に対する平均単語数の分布 – PaulGraham
Fig. 4　The distribution of average term frequencies for each probability on the Paul Graham method.

なく，特徴的な単語のみが抽出できている．

図 5 は，PaulGraham 方式を用いた場合の，spam 度に対するメール数の分布を表している．横軸には，spam 度 $P(M)$ を，縦軸には，spam 度を 0.05 ずつの区間に分けて，各区間に含まれるメール数をとったヒストグラムである．ここで，ham メールについてプロットしたグラフ（Ham mails）は $[0, 0.05]$ の範囲に，spam メールについてプロットしたグラフ（Spam mails）は $(0.95, 1]$ の範囲に多く分布しており，カテゴリ間のマージンは非常に大きい．文献 2) によると，PaulGraham 方式での閾値 t は 0.9 とするとあるが，このグラフからも，その値が妥当であることが確認できる．ただし，spam 度が $(0.95, 1]$ の範囲には，ham メールが含まれており，誤検出を低減するために，閾値 t の範囲を $[0.95, 1)$ とする．

2.5　Robinson 方式

Robinson 方式は，Gary Robinson が PaulGraham 方式を改良した方式[4] である．まず，単語 w_i ごとの spam 確率 $p(w_i)$ を，

$$p(w_i) = \frac{s_i/S}{h_i/H + s_i/S} \tag{6}$$

のようにバイアスをかけずに計算し，尤度 $f(w_i)$ を次式のように計算する．

$$f(w_i) = \frac{s \cdot x + n_i \cdot p(w_i)}{s + n_i} \tag{7}$$

このとき，x は未知語の spam 確率，s は x の予測に与える強さ（strength），n_i は単語 w_i の出現回数（$h_i + s_i$）である．文献 4) に準拠すると，$x = 0.5$，$s = 1$ である．

また，単語 $w_1, \cdots, w_N$ を含むメール M の spam 度 $P(M)$ を次式で表すと，

$$S = 1 - \{\prod_{i=1}^{N}(1 - f(w_i))\}^{\frac{1}{N}},$$

$$H = 1 - \{\prod_{i=1}^{N} f(w_i)\}^{\frac{1}{N}},$$

$$P(M) = \frac{S - H}{S + H} \tag{8}$$

spam 度 $P(M)$ が閾値 t を上回った場合，メール M は spam であると判定できる．このとき，閾値 t については，文献 4) に明記されていないため，2.6 節で分析した結果を用いる．

2.6 Robinson 方式の分析

図 6 は，Robinson 方式を用いた場合の，spam 度に対するメール数の分布を表している．横軸には，spam 度 $P(M)$ を，縦軸には，spam 度を 0.05 ずつの区間に分けて，各区間に含まれるメール数をとったヒストグラムである．ここで，ham メールについてプロットしたグラフ（Ham mails）は $[0, 0.55]$ の範囲に多く分布し，spam メールについてプロットしたグラフ（Spam mails）は，$(0.6, 1]$ の範囲に，多く分布している．この分布状況から，閾値 t の範囲を $[0.55, 0.65]$ とする．

2.7 Robinson-Fisher 方式

Robinson-Fisher 方式は，Thunderbird [5] 等のフリーウェアで採用されている方式として定評がある．単語 w_i ごとの spam 確率 $p(w_i)$ および尤度 $f(w_i)$ を

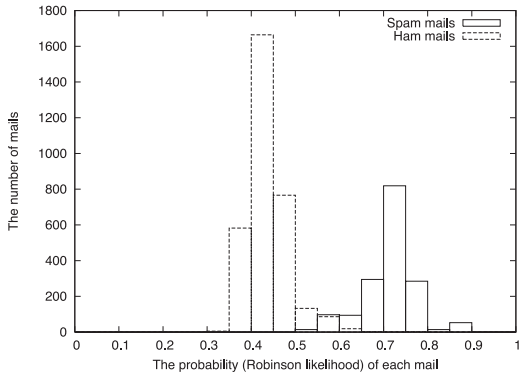

図 6　spam 度に対するメール数の分布 – Robinson
Fig. 6　The distribution of mails for each probability on the Robinson method.

$$p(w_i) = \frac{s_i/S}{h_i/H + s_i/S} \tag{9}$$

$$f(w_i) = \frac{s \cdot x + n_i \cdot p(w_i)}{s + n_i} \tag{10}$$

とする．また，単語 $w_1, \cdots, w_N$ を含むメール M の ham 性 H と spam 性 S から，spam 度 $P(M)$ を次式で表す．

$$H = C^{-1}(-2ln\prod_{i=1}^{N} f(w_i), 2N),$$

$$S = C^{-1}(-2ln\prod_{i=1}^{N}(1 - f(w_i)), 2N),$$

$$P(M) = \frac{1 + H - S}{2} \tag{11}$$

ここで，C^{-1} は逆 χ^2 関数とし，spam 度 $P(M)$ が閾値 t を上回った場合，メール M は spam であると判定する．

ただし，逆 χ^2 関数 C^{-1} は計算コストが大きいため，bsfilter [6] 等が採用している近似方法と同様に，

$$H' = 1 - C(-2ln\prod_{i=1}^{N} f(w_i), 2n),$$

$$S' = 1 - C(-2ln\prod_{i=1}^{N}(1 - f(w_i)), 2n),$$

$$P(M)' = \frac{1 + H' - S'}{2} \tag{12}$$

として，χ^2 関数 C を用いて近似する．

2.8 Robinson-Fisher 方式の分析

図 7 は，Robinson-Fisher 方式を用いた場合の，spam 度に対するメール数の分布を表している．横軸には，spam 度 $P(M)$ を，縦軸には，spam 度を 0.05 ずつの区間に分けて，各区間に含まれるメール数をとったヒストグラムである．ここで，ham メールについてプロットしたグラフ（Ham mails）は $[0, 0.05]$ の範囲に多く分布し，spam メールについてプロットしたグラフ（Spam mails）は $(0.95, 1]$ の範囲に，多く分布している．また，$(0.5, 0.55]$ の

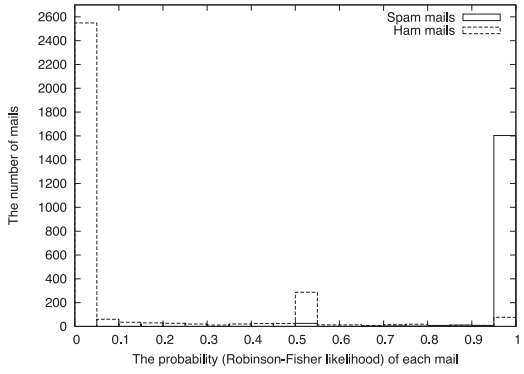

図 7　spam 度に対するメール数の分布 – Robinson-Fisher
Fig. 7　The distribution of mails for each probability on the Robinson-Fisher method.

範囲にも，約 300 通のメールが分布しており，これらのメールを判定不能とすることもできるのが Robinson-Fisher 方式の特徴の 1 つである．ただし本実験では，誤検出率を低減するため，これらのメールを ham メールと判断し，閾値 t の範囲を $[0.55, 1)$ とする．

3. 提 案 方 式

我々はまず，従来方式について，特に誤検出率（FPR）を低減させるよう閾値 t を設定した場合に，より高い精度（TNR）を得るため，各方式のグラフを用いて考察し，課題を述べる．次に，従来方式の課題を改善する方法について検討し，改善方法を提案する．

3.1 従来方式の課題

従来方式で誤検出が起こる主な原因は，各方式の spam 度に対するメールの分布に，ham メールと spam メールが混在する範囲が存在することである．この問題を回避するためには，PaulGraham 方式では，単語の選択方法を改善すること，Robinson 方式および Robinson-Fisher 方式では，事前分布を修正することが考えられる．

特に，PaulGraham 方式では，図 5 を見ると，spam 度が $[0, 0.05]$ または $(0.95, 1]$ の範囲にのみ分布しており，spam 度の曖昧なメールはない．ここで，PaulGraham 方式の閾値 t を 0.9 としたとき，誤検出される ham メールは，spam 度が 0.95 以上であることから，選択された 7 語の単語の spam 確率は，相乗平均により 0.99270（$= 0.95^{(1/7)}$）以上である．このことから，ham メールにもかかわらず，spam 度の計算に利用される単語のほとんどが，spam 確率の高い単語であるといえる．

このように，ham メールにもかかわらず，spam 度の高い単語が多く含まれる原因としては，spam メールのメール送信者が，判定を誤らせることを目的として，以下のような spam メールを作成した場合が考えられる．

CASE 1　spam メールの中に，ham メールに含まれていそうな単語「Yahoo!」「Amazon」「楽天」等を意図的に混入させる．

CASE 2　spam メールと判断されそうな単語を，「Via_g_r_a」「ばいアぐ RA」「VI@GRA」等の単語（未知語）に置き換えて，spam 確率の高い単語として抽出できないように，spam メールの内容を構成する．

CASE 1 の場合，意図的に混入された単語の spam 確率が高くなり，**CASE 2** の場合は，置き換えられた単語以外の単語の ham 確率が低くなる．

このような問題は，従来方式のすべてに共通する課題であるが，特に PaulGraham 方式では，メール内の単語の spam 確率の分布状況とは無関係に，spam 確率が 0.5 から離れている単語を選択するため，意図的に混入された単語や，置き換えられた単語の影響で，spam 確率や ham 確率が偏ると，メールの spam 度を計算するために選択される単語が，spam 確率の高い単語だけになってしまい，メールの spam 度が極端に高い値となる場合がある．

以上の理由から，PaulGraham 方式では，ham メールに含まれている単語の spam 確率に偏りがある場合，その偏りがメールの spam 度を計算するための単語の選択方法にも影響し，メールの spam 度が，$[0, 0.05]$ または $(0.95, 1]$ の範囲に極端に偏る．

3.2 PaulGraham 方式の改良

我々は，従来方式の課題を改善するための 1 つの方法として，PaulGraham 方式をもとに，メールの spam 度を計算するための単語の選択方法を改良することで，誤検出を減らすことを考える．具体的には，1 通のメール内に含まれる単語のうち，ham 確率の高い単語と，spam 確率の高い単語の同数を，単語の spam 確率の大きさ順に選択する（Bipolar 方式と呼ぶ）．

Bipolar 方式は，単語の spam 確率や ham 確率が偏ってしまった場合も，メール内で相対的に高い spam 確率，または相対的に高い ham 確率の単語の同数を選択し，メールの spam 度の計算に利用する．このため，メールの spam 度の極端な偏りは軽減され，閾値 t の調整で，誤検出を低く抑えることができる．

3.3 Bipolar 方式

Bipolar 方式では，まず，ある単語 w_i が spam として登場した回数を s_i，ham として登

場した回数を h_i とし，spam メールの総数を S，ham メールの総数を H としたとき，spam 確率 $p(w_i)$ を次式で表す．

$$p(w_i) = \frac{s_i/S}{h_i/H + s_i/S} \tag{13}$$

また，メール M に含まれる単語 $w_1, \cdots, w_N$ を $p(w_i)$ の大きい順にソートし，spam 確率 $P(S)$ と ham 確率 $P(H)$ を，spam 確率の高い単語 n 語と ham 確率の高い単語 n 語により

$$P(S) = \prod_{i=1}^{n} P(S|w_i) = \prod_{i=1}^{n} p(w_i),$$
$$P(H) = \prod_{i=n'}^{N} P(H|w_i) = \prod_{i=n'}^{N} (1 - p(w_i))$$

と表す．ここで，$n' = N - n + 1$ とし，さらにメール M の spam 度 $P(M)$ を

$$P(M) = \frac{P(S)}{P(S) + P(H)} \tag{14}$$

とすると，spam 度 $P(M)$ が閾値 t を上回った場合に，メール M は spam であると判定できる．なお，閾値 t および n は，3.4 節で分析した結果を用い，単語が未知語の場合は，文献 2) に準拠し，$p(w_i) = 0.4$ とする．

3.4　Bipolar 方式の分析

図 8 は，Bipolar 方式で閾値 t は 0.6 とし，10-folds の交差検定により計測した場合の，単語数 n と精度（TNR）の関係を表す．この結果，単語数 n が 10 のとき 0.98317 と，最も高い精度のため，$n = 10$ とする．

図 9 は，Bipolar 方式を用いた場合の，spam 確率に対する選択された平均単語数の分布を表している．横軸には，spam 確率 $p(w_i)$ を，縦軸には，spam 確率を 0.05 ずつの区間に分けて，各区間に含まれる平均単語数（単語数 ÷ メール数）をとったヒストグラムである．ここで，spam 確率が $(0.9, 1]$ の範囲には，ham メールにも spam メールにも含まれる単語が多く，spam 確率が $[0, 0.1]$ の範囲には，ham メールに含まれる単語が多く分布し，特徴的な単語を抽出できる．このとき，spam 確率が $(0.05, 0.4]$ の範囲にも，単語が分布しており，PaulGraham 方式では単語の分布がなかった $[0.1, 0.9]$ の範囲に，単語の分布があり，単語選択に極端な偏りがないことを裏付ける．

図 10 は，Bipolar 方式を用いた場合の，spam 度に対するメール数の分布を表している．

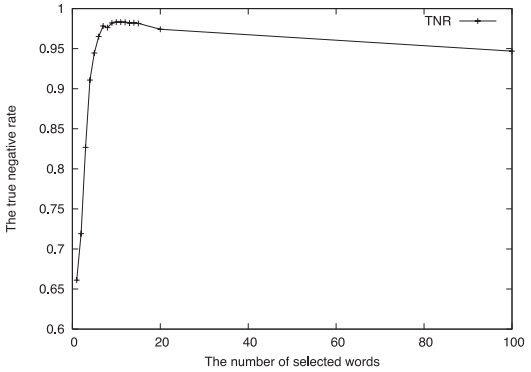

図 8　選択する単語数（n）に対する精度（TNR）— Bipolar
Fig. 8　The true negative rate for the number of selected words on the Bipolar method.

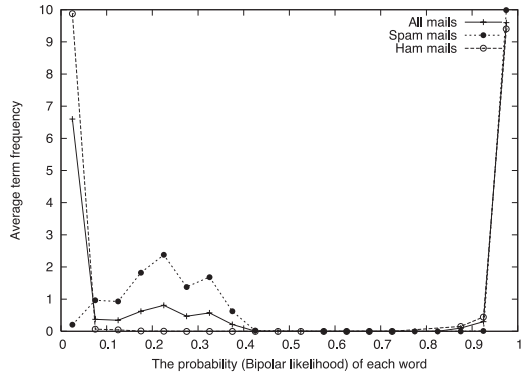

図 9　spam 確率に対する平均単語数の分布 – Bipolar
Fig. 9　The distribution of average term frequencies for each probability on the Bipolar method.

横軸には，Bipolar 方式に基づいて計算した spam 度 $P(M)$ を，縦軸には，spam 度を 0.05 ずつの区間に分けて，各区間に含まれるメール数をとったヒストグラムである．ここで，ham メールについてプロットしたグラフ（Ham mails）は $[0, 0.6]$ の範囲に多く分布し，spam メールについてプロットしたグラフ（Spam mails）は $(0.65, 1]$ の範囲に，多く分布している．この分布状況から，閾値 t の範囲を $[0.6, 0.7]$ とする．

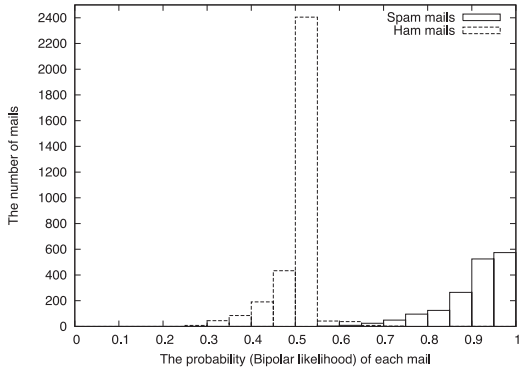

図 10　spam 度に対するメール数の分布 – Bipolar
Fig. 10　The distribution of mails for each probability on the Bipolar method.

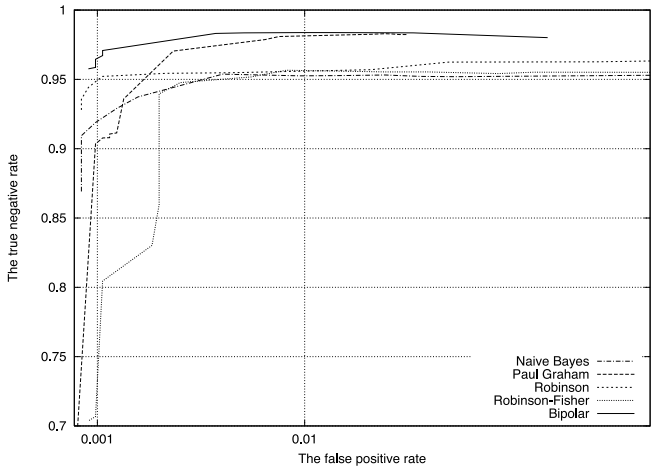

図 11　精度–誤検出率曲線
Fig. 11　TNR-FPR curve of five methods.

4. 実 験 結 果

4.1 交差検定の結果

図 11 は，1.2 節で説明したデータセットを用いて，各方式について，閾値 t を変化させて，10-folds の交差検定により計測した精度（TNR）と誤検出率（FPR）の関係を，曲線で表したグラフである．誤検出率（FPR）の値が小さいとき，精度（TNR）の値がより大きくなる spam メールフィルタが望ましい．このグラフでは，より X 座標の値が小さく，より Y 座標の値が大きい点を通るような曲線は Bipolar 方式である．

4.2 誤検出率を重視

次に，誤検出率（FPR）を一定の値以下とした場合のそれぞれの精度（TNR）を調べる．**表 1** は，各方式の誤検出率（FPR）が 0.01 以下になるよう閾値 t を設定し，10-folds の交差検定により計測した場合の精度および誤検出率である．この条件では，Bipolar 方式の精度が最も高く，次いで PaulGraham 方式の精度が高い結果となった．

表 2 は，各方式の誤検出率が 0.001 以下になるよう閾値 t を設定し，10-folds の交差検定により計測した場合の精度および誤検出率である．この条件でも，Bipolar 方式の精度が最も高く，次いで Robinson 方式の精度が高い結果となった．

以上の結果から，Bipolar 方式は，高い精度を保ちながらも，従来方式よりも誤検出率を低減させることができることを確認した．

表 1　誤検出率 0.01 以下の場合の交差検定の結果
Table 1　10-folds cross validation ($FPR \leq 0.01$).

Method	Threshold(t)	FPR	TNR
Bipolar	0.58	0.00982	**0.98378**
PaulGraham	0.99	0.00890	0.98094
Robinson-Fisher	0.99	0.00920	0.95660
Robinson	0.59	0.00675	0.95478
NaiveBayes	0.59	0.00982	0.95255

TNR means the true negative rate with 10-folds cross validation for each method.
FPR means the false positive rate.

表 2　誤検出率 0.001 以下の場合の交差検定の結果
Table 2　10-folds cross validation ($FPR \leq 0.001$).

Method	Threshold(t)	FPR	TNR
Bipolar	0.73	0.00092	**0.96451**
Robinson	0.62	0.00092	0.94829
NaiveBayes	0.66	0.00092	0.91888
PaulGraham	$[A]$	0.00061	0.90834
Robinson-Fisher	$[B]$	0.00092	0.70614

TNR means the true negative rate with 10-folds cross validation for each method.
FPR means the false positive rate. ($[A] = 1 - 1.6 * 10^{-14}$, $[B] = 1 - 1.6 * 10^{-18}$)

5. おわりに

本稿では，追加学習が高速で，誤検出が十分に少ない spam メールフィルタの開発を目指して，ベイジアンフィルタの 1 つである PaulGraham 方式の単語の選択方法に改良を加え，Bipolar 方式を提案した．

Bipolar 方式の特徴をまとめる．

- ham 確率の高い単語と，spam 確率の高い単語の同数を選択し，メールの spam 度を計算するため，spam 性と ham 性のどちらにも寄与しない単語は，除去される．
- メール内の単語の分布から，相対的に spam 確率の高い単語と，ham 確率の高い単語を選択するため，単語の spam 確率に偏りがある場合にも，メールの spam 度が極端に偏らない．

Bipolar 方式は，ベイジアンフィルタの特性の 1 つである学習および判定の高速性を維持しつつ，誤検出率（FPR）を一定の値以下に保った場合の精度（TNR）が従来方式を上回る．また，図 11 を見ると分かるように，より低い誤検出率を得るために閾値 t を調整しても，精度が下がりにくい．このため，同程度の精度の場合に，従来方式よりも誤検出が少ない spam メールフィルタを実現できた．

ただし，本稿では，メールがすべて自然文で記述されていることを前提としているため，HTML 本文のメールや画像メール，PDF メールといった spam メールに対してはなんら対策を施していない．また，実際の spam メールは日々進化している[7] ことや，ユーザごとに spam メールと判断する基準が異なることもあり，ユーザのフィードバックに対する学習機能には，再現性（reproducibility; 学習効果の定着性：素直さ），感度（sensitivity; 1 回の学習による波及効果の現れやすさ），安定性（stability; 反対学習による誤判定の振れ幅の小ささ）等の観点に基づいて，閾値 t 等のパラメータを動的に調整できることも重要になると考える．

以上のような理由から，今後はさらに，ヘッダ情報や，添付ファイルであるマルチメディアデータを対象とした特徴抽出方法について検討することで，さらなる精度向上を目指す．また，ユーザインタラクションの観点から，学習の特性について検討し，さらなる実用性の向上を実現したい．

参 考 文 献

1) Duda, R.O., Hart, P.E. and Stork, D.G.: *Pattern Classification*, *Second Edition*, pp.61–62, John Wiley & Sons Inc (2000).
2) Graham, P.: A Plan for Spam (2002). http://paulgraham.com/spam.html
3) Graham, P.: Better Bayesian Filtering (2003). http://www.paulgraham.com/better.html
4) Robinson, G.: A Statistical Approach to the Spam Problem, *Linux Journal*, Vol.107 (2003).
5) Thunderbird, Mozilla Foundation. http://www.mozilla.com/en-US/thunderbird/
6) Nabeya, K.: bsfilter. http://bsfilter.org/index-e.html
7) The State of Spam, A Monthly Report, Symantec Corporation. http://www.symantec.com/enterprise/security_response/weblog/security_response_blog/spam/
8) 麻生英樹，津田宏治，村田　昇：パターン認識と学習の統計学新しい概念と手法，統計科学のフロンティア 6，Chapter 3.8，岩波書店 (2003).
9) 特集 spam メールの現状と対策の動向，*IPSJ Magazine*, Vol.46, No.7, pp.739–791, 情報処理学会 (July 2005).
10) 渡部綾太，愛甲健二：スパムメールの教科書，pp.106–114, DATA HOUSE (2006).

(平成 19 年 11 月 22 日受付)
(平成 20 年 1 月 14 日再受付)
(平成 20 年 2 月 4 日採録)

谷岡　広樹

平成 9 年千葉大学工学部電子電子工学科卒業．同年（株）ジャストシステム入社．自然言語処理，情報検索，機械学習等の研究開発に従事．平成 16 年信州大学大学院工学系研究科情報工学専攻博士前期課程修了．平成 20 年信州大学大学院総合工学系研究科システム開発工学専攻博士後期課程修了見込，電子情報通信学会，人工知能学会，ACM，IEEE 各会員．

中川　　尚

平成 9 年京都大学理学部卒業．平成 12 年京都大学大学院理学研究科生物物理修士課程修了．同年（株）ジャストシステム入社．平成 19 年より（株）ジャストシステム福岡研究所勤務．自然言語処理，情報検索，機械学習等の研究開発に従事．

丸山　　稔

昭和 57 年東京大学工学部計数工学科卒業．同年三菱電機（株）入社，先端技術総合研究所勤務．平成 2〜3 年マサチューセッツ工科大学人工知能研究所客員研究員．平成 8 年より信州大学工学部情報工学科准教授．博士（工学）．コンピュータビジョン，学習等の研究に従事．電子情報通信学会，ACM，IEEE 各会員．