社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2009-IOT-4 (36)
2009 / 3 / 6

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

小川 健司† 稲葉 宏幸†

† 京都工芸繊維大学 大学院工芸科学研究科
〒 606-8585, 京都市左京区松ヶ崎御所海道町
E-mail: †{ogawa,inaba}@ice.is.kit.ac.jp

あらまし 近年，パソコンや携帯電話が普及する中，通信手段として電子メールが多く利用されている．その中で，ユーザの意思に関わらず，有害かつ悪質なメールを受信することが多くある．なかには出会い系サイトへの勧誘等の犯罪性が高いメール等もあり，無視できなくなってきた．この対策手段の 1 つとして，フィルタリングがある．特に，ベイジアンスパムフィルタは統計的手法によりメールのスパム確率，つまり迷惑メールである確率を求め，継続的な学習によりフィルタの性能を向上させることができるため幅広く用いられている．しかし，ベイジアンスパムフィルタでも検知が難しいメールが存在する．このようなメールはメール本文中に含まれる単語の間に☆や★などの記号を挟んだり，記号を羅列している傾向がある．

本報告では，まず最初に最近の迷惑メールと正規メール各 1000 通ずつについてメール本文中の記号と未知語の分布を調査した結果を示す．その結果，両者の間には明確な分布の違いがあることが明らかになった．そこでその違いをベイジアンスパムフィルタにおけるスパム確率の算出の際に利用する新たなフィルタを提案し，その性能を評価する．
キーワード 迷惑メール, 記号, 未知語, ベイジアンフィルタ, フィルタリング

Proposal on Bayesian Spam Filter Using Distribution of Symbol and Unknown Word

Kenji OGAWA† and Hiroyuki INABA†

† Kyoto Institute of Technology
Goshokaidoucho, Matsugasaki, Sakyo-ku, Kyoto 606-8585 JAPAN
E-mail: †{ogawa,inaba}@ice.is.kit.ac.jp

Abstract Recently, spam mail, that is an irrelevant and unsolicited mail, is one of the most serious problem in Internet. A Bayesian spam filter is a popular method to deal with the problem at a recipient computer. However, a mail which includes many symbols and unknown words is hardly classified accurately by a conventional Bayesian spam filter. In this report, we propose a new Bayesian type spam filter which utilize a distribution of symbols and unknown words included in the received mail. We confirm the performance of the proposed method by experiment.
Key words UBE, unknown word, symbol, Bayesian filter, Filtering

1. まえがき

近年，パソコンや携帯電話の普及により，手軽な情報通信手段のひとつとして電子メールが広く利用されている．その一方で，インターネットの匿名性の高さを利用して，これを悪用した有害な電子メールの流通が社会問題となっている．

このような電子メールは，ユーザの意思に関わらず一方的に送られて来るもので，いわゆる「迷惑メール」と呼ばれている．これらには，出会い系サイトの登録の勧誘やワンクリック詐欺等を目的とした犯罪性の高いメールも含まれている．これらは近年増加傾向にあり，深刻な被害が出ているケースもあり，無視できなくなってきた．

迷惑メールの問題を解決するために，様々な対策技術が開発され，導入されている．しかし，これらの対策技術には長所と短所があり，いくつかの対策技術を組み合わせることで，受信する迷惑メールの量を大幅に減らすことができるが，それでも 100 %全ての迷惑メールを受信拒否することはできない．

日本語の迷惑メールをクライアントで検知する際に特に問題となるのが，日本語特有の表記方法に関する問題である．一般に，単語が空白で区切られている英文と違い，区切りのない

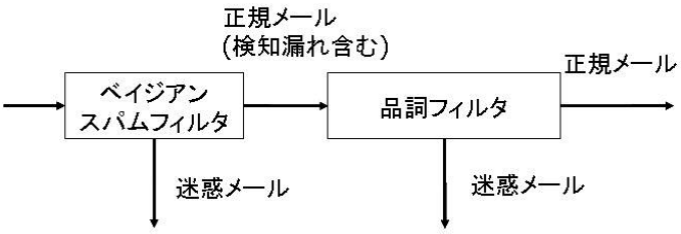

図 1　文献 [1] での提案方式の処理の流れ
Fig. 1　A processing flow of the method proposed in Report [1].

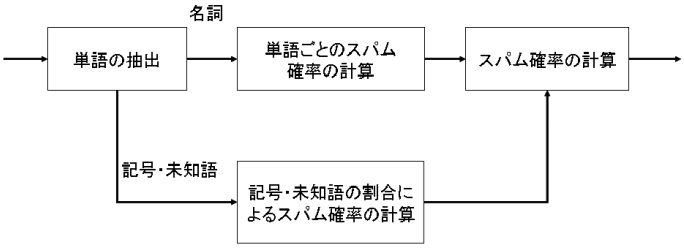

図 2　提案方式の処理の流れ
Fig. 2　A processing flow of proposal method.

日本語は，単語ごとに抽出することが困難である．さらに迷惑メールによく使われる単語を，その単語中に記号や伏字を挟んだり，代用するなどしてごまかすことによって，単語として抽出しにくいように書かれた迷惑メールが問題になっている．実際，文献 [1] では，日本語の迷惑メールに含まれる品詞情報は記号・未知語の割合が多いことが報告されており，従来のベイジアンスパムフィルタ [2] では性能が劣化することがある．この対策として，文献 [1] では，ベイジアンスパムフィルタによる判定後に，品詞情報によるフィルタリングを行うことで，このような迷惑メールを検知することを試みているが，最近は迷惑メールの中にも記号・未知語の割合が低いものや，正規メールでも広告メールなどでは記号の割合が多いものが増えており，必ずしもうまくフィルタリングできるとは限らなくなっている．

本報告ではこのような迷惑メールへの対策として，ベイジアンスパムフィルタの評価指標に，記号と未知語の割合を評価手法として取り入れた，新たなフィルタリング方法を提案し，その評価を行う．

2. 従来方式の問題点

文献 [1] では，図 1 のように，受信メールのうち 1 次フィルタのベイジアンスパムフィルタで正規メールと判断されたメールは，さらに 2 次フィルタである，品詞フィルタを通る流れになっている．

しかし，この方法では，2 次フィルタでの判定の際に，記号と未知語の割合のみが考慮されるため，品詞情報が迷惑メールに類似している正規メールは誤検知されてしまうことがある．このような正規メールとして，テキスト整形をしたメール，プロバイダからのお知らせメール，Web サービスの利用情報や登録情報が記載されたメール等が挙げられる．この点を改良するため，以下では新たなスパムフィルタを提案する．

3. 提案方式の概要

3.1 提案方式の処理の流れ

ここではベイジアンスパムフィルタにおける判定の際に，記号と未知語の割合を指標の 1 つとして用いたフィルタを提案する．なお，本研究ではベイジアンスパムフィルタの実装として日本語にも対応している bsfilter [2] を，単語抽出には MeCab [3] をそれぞれ利用している．

まず，従来のベイジアンスパムフィルタの処理の流れを述べる．一般に，ベイジアンスパムフィルタ（Gary-Robinson 方式 [4]）におけるスパム判定の処理は，まず単語抽出をし，その際に名詞もしくはコーパスにない単語である未知語と推定されたものに対し，その出現回数を用いて，式 (1) により，単語 w_i に対して $p(w_i)$ を計算する．

$$p(w_i) = \frac{(\frac{b_i}{n_{bad}})}{(\frac{g_i}{n_{good}}) + (\frac{b_i}{n_{bad}})} \tag{1}$$

g_i：非スパムコーパス中に w_i が出現した回数
b_i：スパムコーパス中に w_i が出現した回数
n_{good}：非コーパス中のメールの総数
n_{bad}：スパムコーパス中のメールの総数

この $p(w_i)$ を用いて，名詞スパム確率 $f(w_i)$ を次のように計算する．

$$f(w_i) = \frac{s \cdot x + n \cdot p(w_i)}{s + n} \tag{2}$$

x：今まで 1 度もメールの中に出現していない単語がはじめてメール中に出現したときに，そのメールが迷惑メールである予測確率
s：x の予測に与える強さ
n：単語 w_i の出現回数

式 (2) において，x と s の値は，$x = 0.5$，$s = 1$ が妥当であるとされている．

次に，この $f(w_i)$ を用い，P，Q を次式により計算する．

$$P = 1 - \left(\prod_{i=1}^{n} (1 - f(w_i)) \right)^{\frac{1}{n}} \tag{3}$$

$$Q = 1 - \left(\prod_{i=1}^{n} f(w_i) \right)^{\frac{1}{n}} \tag{4}$$

最後に，以下の式 (5) によって P と Q からメールのスパム確率 S を求め，最終的な判定を出力する．

$$S = \frac{P - Q}{P + Q} \tag{5}$$

文献 [2] では，$S = 0.582$ をデフォルトとしている．

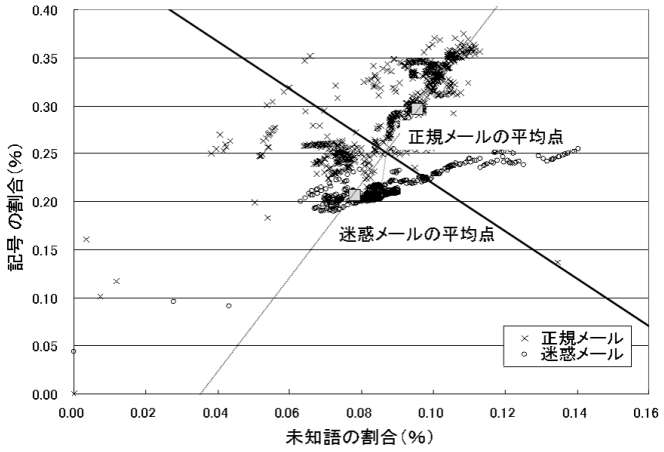

図 3　メール各 1000 件ずつの記号と未知語の割合による分布
Fig. 3　Distribution of spam and ham mail in each 1000 mails.

一方，提案方式では，図 2 のような処理の流れになっている．まず，形態素解析を行い，単語抽出をする．その際に，名詞と判定された単語は従来のベイジアンスパムフィルタと同様に，名詞スパム確率（式 (2)）を求める．記号または未知語と判定された単語は，メール本文中の記号と未知語の割合によって品詞スパム確率を求め（3. 2 で詳述），それら 2 つの値を用いて，最終的にメールのスパム確率を求め（3. 3 で詳述），判定結果を出力する流れになっている．

3. 2　記号と未知語の処理

著者宛てに，2008 年 1 月から 2008 年 4 月までの間に届いた正規メールと迷惑メール各 1000 件ずつの記号と未知語の分布を調べると，図 3 のようになっていることが明らかとなった．

文献 [1] に示されている分布とは形状が異なるが，正規メールと迷惑メールの間に分布の違いがあることは同様である．これを各分布の平均点を結ぶ直線の上に各点を垂直に射影すると，図 4 のようになる．図 4 における，各平均点付近の正規メール，迷惑メールの代表的な例を，図 5 および図 6 に示す．さらに，両分布の平均点の中点 m(図 4 では-0.05264 の点に相当) を閾値として設定し，以下のように品詞スパム確率 f_{info} を定めた．

（1）　射影後の品詞の割合 ≧ m のとき

$$f_{info} = 1 \tag{6}$$

（2）　射影後の品詞の割合 ＜ m のとき

$$f_{info} = 1 \times 10^{-8} \tag{7}$$

このように，メール毎に記号と未知語の割合によって品詞スパム確率を定め，記号と未知語以外の名詞のスパム確率と合わせ，次節の方法によりメールのスパム確率を求める．

3. 3　スパム確率の計算

記号と未知語以外の名詞によって求めた名詞スパム確率と，式 (6) または式 (7) によって，記号と未知語の割合によって求めた品詞スパム確率を合わせて，メールスパム確率 S を算出する．実際には，式（3）および式（4）を以下のようにそれぞれ

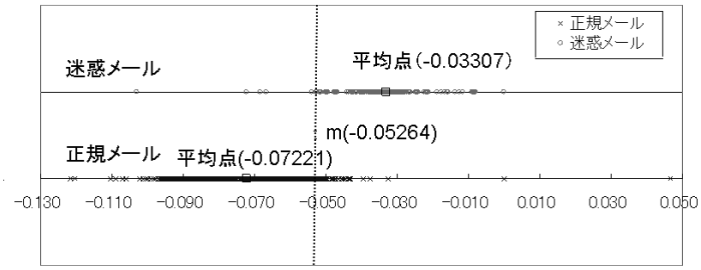

図 4　射影後の記号と未知語の割合による分布
Fig. 4　Distribution of symbol and unknown word after projection.

忙しいIT担当者のためのDailyマガジン ... 興味分野の情報を毎日お届け

キーマンズネット | Keyレポ ★ Daily < 2008/06/04 >

C | O | N | T | E | N | T | S |

■■本日の製品情報■■

1:【ID管理ツール】製品選定に迷っているアナタに、選び方を伝授！...PR
→ http://www.keyman.or.jp/3w/prd/23/10011523/?ml=d20080604

2:【簡単BI】ドラッグ＆ドロップだけで、業務データが分析できる！...PR
→ http://www.keyman.or.jp/3w/prd/78/10011278/?ml=d20080604

3:オラクルの最新SaaS型CRMを活用して、新しい顧客管理の実現へ！...PR
→ http://www.keyman.or.jp/3w/prd/81/10011181/?ml=d20080604

4:100社30万クライアントの販売実績！ IT戦略に、ManagementCore...Keyレポ
→ http://www.keyman.or.jp/3w/prd/54/10010754/?ml=d20080604

5:使いやすさを標準装備！直感的に扱えるクライアント管理ツール！...Keyレポ
→ http://www.keyman.or.jp/3w/prd/29/10010729/?ml=d20080604

※お届けしている情報が今欲しい情報と違う場合、興味分野を変更ください
→ http://www.keyman.or.jp/3w/prd/00/03006000/?ml=d

1:【ID管理ツール】製品選定に迷っているアナタに、選び方を伝授！...PR

→ http://www.keyman.or.jp/3w/prd/23/10011523/?ml=d20080604

■カテゴリ...PKI・暗号化・認証
■内　容...Secured AccountOne
■会社名...ソフトバンク・テクノロジー

⇒アカウント管理の膨大な手間/コストの軽減、情報漏洩防止、日本版SOX法対応などを理由に、ID管理ツールの導入を検討している企業は多いです。しかし...
『製品が似ていて、どれがいいのかわからない！』という人も多いでしょう。

図 5　正規メールの例
Fig. 5　Example of ham mail.

★★★★★★★★★★★★★★★★★★★★★★★★★★★★★★★★★★★★

lnn
(|口 ├┼げ ん 玉 | げん玉電鉄 結果発表！(●˝▽˝●)／
[三三三L | ======== |
∠7◎◎◎=◎ ◎ ◎ ◎ by げん玉(リアルワールド)

★★★★★★★★★★★★★★★★★★★★★★★★★★★★★★★★★★★★

2008年9月8日(月)14:00 ～ 2008年9月15日(月)13:59
のげん玉電鉄ゴールランキングを発表します！

☆──────────────────────☆
| <チャンピオン> |
| 賞 金 :100,000ポイント獲得 おめでとうございます！ |
☆──────────────────────☆

チャンピオン:こうぼ様

☆──────────────────────☆
| <2位～10位> |
| 賞 金 :1,000ポイント獲得 おめでとうございます！ |
☆──────────────────────☆

2位:luftballon様
3位:宇宙最弱様
4位:まー君様
5位:花火様
6位:シャケデラ様
7位:kgons様
8位:みのごんた様
9位:フィーゴ様
10位:monaco様

☆──────────────────────☆
| <切り番当選者> |
| 賞 金 :1,000ポイント獲得 おめでとうございます！ |
☆──────────────────────☆

図 6　迷惑メールの例
Fig. 6　Example of spam mail.

変更する．

$$P = 1 - (1 - f_{info})^{\frac{1}{n}} \cdot \left(\prod_{i=1}^{n} (1 - f(w_i)) \right)^{\frac{1}{n}} \quad (8)$$

$$Q = 1 - (f_{info})^{\frac{1}{n}} \cdot \left(\prod_{i=1}^{n} f(w_i) \right)^{\frac{1}{n}} \quad (9)$$

最終的にメールスパム確率 S を求める式は，Robinson 方式と同じである．

$$S = \frac{P - Q}{P + Q} \quad (10)$$

これは，品詞スパム確率を名詞スパム確率の 1 つとして扱い，スパム確率を計算することを意味している．

4. シミュレーションによる提案方式の評価

4.1 シミュレーション条件

3. で提案したベイジアンスパムフィルタの性能を評価するため，以下の条件を設定して，検知シミュレーション実験を行う．

- 評価対象:
  提案方式および従来方式 (bsfilter)

- 評価項目:
  FPR(False Positive Rate) と FNR(False Negative Rate) による ROC 曲線 (Receiver Operating Characteristic)

- 評価条件
  学習用メール数: 正規メール，迷惑メール　各 1000 件ずつ
  評価用メール数: 学習用メールとは異なる正規メール，迷惑メール各 1000 件ずつ
  評価対象範囲：メール本文
  単語抽出法: 形態素解析 (MeCab)
  スパム確率評価式: 提案方式：3.3 で示した式
  従来方式：Gary Robinson 方式

4.2 実 験 結 果

実験結果の ROC 曲線を図 7 に示す．ここでの ROC 曲線とは，スパム確率 S に対する閾値を変化させることにより，描いたものである．従来方式よりも性能が改善されていることがわかる．一般に，FPR（正規メールが迷惑メールと間違って判定される割合）は，3 %〜10 %程度に設定されることが多いと考えられるが，提案方式ではこの範囲において，FNR が，従来方式に比べ約 13 %程度改善されていることが明らかになった．

本実験では，未知語と記号の分布の違いをスパム判定に取り入れた効果を明らかにするために，メール本文のデータのみを使用している．実際にベイジアンスパムフィルタを．用いる場合は，メールヘッダに含まれる各種情報もスパム判定に用いることが一般的である．提案方式の場合も，メールヘッダの情報もあわせて用いることにより，FPR，FNR 共に更に減少することが期待できる．

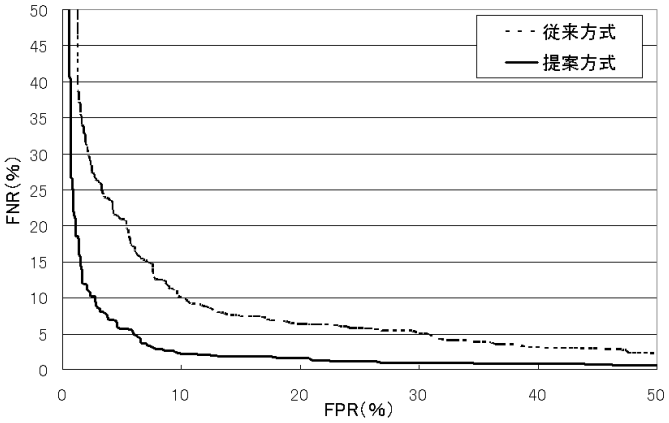

図 7　従来方式と提案方式の ROC 曲線
Fig. 7　ROC curve of both proposal method and the conventional method.

5. む す び

本報告では，スパム確率の高い単語に記号等を混ぜる等，従来のフィルタリング技術では検知が難しいメールにも対応可能な新たな対策法を提案した．提案手法は，正規メールと迷惑メールについて，記号と未知語の割合の分布が異なることを利用し，それらをメールスパム確率を求める際の指標の一つとして用いる方式である．提案方式を実装し，実際に受信したメールを用いて実験を行ったところ，本方式は，従来方式と比べ，良好な結果を得ることが確認でき，記号や未知語が多く含まれる最近のメールの傾向に対応できることが明らかになった．
提案方式で利用している記号と未知語の割合の分布は，環境や時間経過によって変化する可能性があるため，今後，品詞フィルタのパラメータを受信データから動的に学習していく方法等も検討する必要があろう．

文　　献

[1] 近藤智司, 西岡悠, 稲葉宏幸，”品詞情報に着目した日本語迷惑メールフィルタリングの提案”，電子学会 電子・情報・システム部門大会講演論文集 (CDROM), MC2-6,pp.47, 2007.

[2] ”bsfilter / bayesian spam filter / ベイジアンスパムフィルタ”, http://bsfilter.org/, 2009/1 現在

[3] ”MeCab: Yet Another Part-of-Speech and Morphological Analyzer”, http://mecab.sourceforge.net/, 2009/1 現在

[4] Gary Robinson, ”Spam Detection”, http://radio,weblogs.com/0101454/stories/2002/09/16/spamDetection.html, 2009/1 現在

[5] Richard O.Duda, Peter E.Hart, and David G.Stork, ”Pattern Classification”, 2nd ed, 2000, 2009/1 現在

[6] Paul Graham, ”スパムへの対策 — A Plan for Spam”, http://practical-scheme.net/trans/spam-j.html, 2009/1 現在

[7] Paul Graham, ”ベイジアンフィルタの改善 — Better Bayesian Filtering”, http://practical-scheme.net/trans/better-j.html, 2009/1 現在