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

2008－ＤＰＳ－134 (49)
2008－CSEC－ 40 (49)
2008／3／7

発信元情報を適用した ベイジアンスパムフィルタ方式の提案

伊藤 朋哉† 寺田 真敏†† 土居 範久†

† 中央大学大学院 理工学研究科 情報工学専攻 〒112-8551 東京都文京区春日 1-13-27
†† 中央大学研究開発機構
〒112-8551 東京都文京区春日 1-13-27 中央大学後楽園キャンパス 3 号館 12 階

あらまし 今日，電子メールの普及とともに，営利目的のメールを大量配信するスパムメールが社会問題となっている．技術的な対策のひとつであるベイジアンフィルタはメール本文の単語の特徴に対して有効であるが，スパムメール送信者により回避のための細工がおこなわれていることも多い．そこで本稿では，ベイジアンフィルタの判定を補完する方法として，送信元の情報から踏み台にされているサーバが送信元となっているなどといったスパムメールに特徴的な傾向をベイジアンフィルタに学習させ，スパムメール判定に利用する手法を提案する．提案手法のシステムを実装し，筆者が実際に受信したメールを利用した評価の結果から提案手法の有効性を示す．

Proposal of a Bayesian Spam Filter Method by Applying Originator Information

Tomoya Ito† Masato Terada†† Norihisa Doi†

† Graduate School of Science Engineering, Chuo University.
1-13-27 Kasuga, Bunkyo-ku Tokyo 112-8551, Japan
†† Research and Development Initiative Chuo University.
12th Floor, Chuo University Korakuen Campus,
1-13-27 Kasuga, Bunkyo-ku Tokyo 112-8551, Japan

Abstract Spam mail which is unwanted e-mail with commercial content comprises 84% of all the e-mail in the world in 2007. A large amount of spam mail occupies network traffics and transfer delays occur. The Bayesian Spam Filter is effective way to filter spam mail, but it relays on words included e-mail and there are some methods for avoiding the Bayesian Filter by spam mail senders. To detect the mails which are not detected by the Bayesian Filter, we propose the method to learn extra originator information to the Bayesian Filter and to judge e-mails with spam mail characteristics included the information. We evaluated effectiveness of the proposed method by judging e-mail which is sent to author with this proposed method.

1. はじめに

今日インターネットの普及とともに営利目的のメールを無差別に大量配信するスパムメールが社会問題となっている．すべての電子メールに対するスパムメールが占める割合は近年高い状態が続いており，2007 年は 84%を占めているという調査結果が存在する[1]．スパムメールは大量のメール送信によりネットワーク通信の負荷が大きくなるため転送遅延が起きてしまうなどの問題があり，技術的な対策が求められている．

スパムメールに対する技術的な対策のひとつとしてベイジアンフィルタ[2]が注目されている．ベイジアンフィルタは過去に受信したメールの情報から，判定対象のメールがスパムメールである確率を計算する方法である．ベイジアンフィルタでは，まずスパムメールと正規メールの学習をおこなう．学習ではメール本文中に含まれる単語についてスパムメールおよび正規メールそれぞれの出現頻度よりスパムメールらしさの確率を計算する．その後，判定時には，対象となるメール本文中に含まれている単語を学習したデータを参照し，そのメールのスパムメールらしさを算出する．

しかし近年のスパムメールは，無意味な単語を挿入するといったベイジアンフィルタを回避するための細工を施したメールが多く，その影響により検出の精度が落ちてしまう問題がある．

そこで本研究では，スパムメールに特徴的な傾向が表れる送信元の情報を新たにベイジアンフィルタに学習し，スパムメール判定に利用する方法を提案する．本研究でベイジアンフィルタに新たに学習・判定させる情報は次の 3 つである．

(1) 直前の送信元 IP アドレスから，メール送信元となるサーバの情報

(2) ドメインと送信者認証の結果から，ドメインの偽装の有無の情報

(3) 時差表記と IP アドレスの情報から，メール作成国とメール送信国が一致しない組み合わせの特徴の情報

これらの情報はスパムメールに特徴的な傾向が現れ，送信者による改変が難しいものである．そのためこの特徴を利用することで，ベイジアンフィルタで判定ができなかったメールに対して，補完的な判定をおこなうことができる．

提案方式を実装し，実際に筆者が受信したスパムメールを利用した評価により提案手法の有効性を示す．

2. ベイジアンフィルタ

2.1 ベイジアンフィルタの概要

ベイジアンフィルタは過去に受信したメール本文中の単語を学習し，判定対象とするメールのスパムらしさの確率を算出する方法である．次の手順でベイジアンフィルタによるスパムメールの判別をおこなう．

(1) 学習
過去に受信したスパムメールおよび正規メールの本文の文章の単語を抽出する．単語それぞれのスパムメールおよび正規メールに出現した回数からスパムメールらしさを示す確率を算出し学習データベースに記録する．

(2) 判定
学習時と同様，対象となるメールの文章から単語を抽出する．学習したデータベースから，抽出した単語のスパムらしさを検索し，その値の結合計算によって判定するメールのスパムらしさを導き出す．

ベイジアンフィルタにより算出されるスパムらしさは **0** から **1** までの値で記され，**0** に近いほど正規メールらしいことを示し，**1** に近いほどスパムメールらしいことを示す．

2.2 スパムメール判定の際の問題点

ベイジアンフィルタをスパムメール判定に利用する際の問題点として，判定結果がメール文中の単語に依存してしまい，スパム送信者によりベイジアンフィルタを回避するためのかく乱行為がおこなわれやすいことがあげられる．次にスパム送信者がおこなうかく乱行為の例をあげる．

- 短い文章の使用
- 文章の代わりに画像を使用
- 無意味な文章の挿入
- 意味が通るように単語を変換

例）WATCH → W4TCH

このような送信者のかく乱行為により，ベイジアンフィルタの判定精度が低下してしまうことがある．そこで本研究ではスパムメールに特徴的な傾向が表れる送信元の情報をベイジアンフィルタで学習し，判定に利用する方法について提案する

3. 提案方式

3.1 提案の概要

本研究ではベイジアンフィルタの判定精度の向上を目的に，スパムメールに特徴的な傾向が表れる送信元に関する情報をメールから得て，その情報をベイジアンフィルタにより学習し，判定に利用する方法を提案する．またベイジアンフィルタを用いた学習や判定がおこなえるように，得られた情報を文字列に変換する作業をおこなう．

この方法によって得られる情報は，単語の特徴には依存しないため，これまでベイジアンフィルタで判定できなかったメールに対して補完的な判定ができる．

次節から本提案方式でスパムメールの判定に用いる情報を得るための検査項目の概要について詳細を述べる．

3.2 提案する検査項目の概要

3.2.1 送信元の IP アドレス

本項目は，信頼できるメールサーバが記録した送信元サーバの IP アドレスの情報を学習する．これにより踏み台といわれる第三者にコントロール権限を奪われたサーバを中継することが多いというスパムメールの特徴を判定に用いる．

踏み台となっているサーバはスパムメール送信者によって複数回利用されることがある．その際にメールの内容が同じであるとは限らないため，メール内の文章をもとに判定をおこなうベイジアンフィルタよりも，送信元の特徴を学習する本項目での判定の効果は高いと考えられる．

送信元のサーバを識別する方法としてインターネットサービスプロバイダなど信頼できる受信メールサーバが記録した IP アドレスを利用する．送信元サーバの IP アドレスの取得の概要図を図 **3.1** に示す．

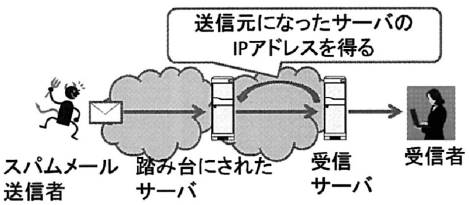

図 3.1 送信元サーバの IP アドレス取得

3.2.2 送信元ドメインの偽装の有無

本項目では，メールに記載されている送信元ドメイ

ンとドメイン偽装の有無の結果を組み合わせて学習する．これにより，ドメインを詐称することが多いというスパムメールの特徴を判定に用いる．

メールを送信するプロトコルである SMTP（Simple Mail Transfer Protocol）[3]において，正しい送信者のドメイン名を宣言することなくメールを送ることができるという構造上の問題がある．スパムメールはこの点を悪用し，有名プロバイダのドメインを詐称するものが多い．この特徴をベイジアンフィルタによる判定に利用する．

ドメインの偽装の有無に関しては SPF（Sender Policy Framework)[4]送信認証技術を利用する．SPF 送信認証はメール送信をしたマシンの IP アドレスからその IP アドレスを管理する DNS サーバにアクセスし，メールに記載されているドメインや IP アドレスの照合をおこなうシステムである．SPF 送信者認証技術の概要図を図 3.2 に示す．また問い合わせにおける返答とその概要を表 3.1 にまとめる．

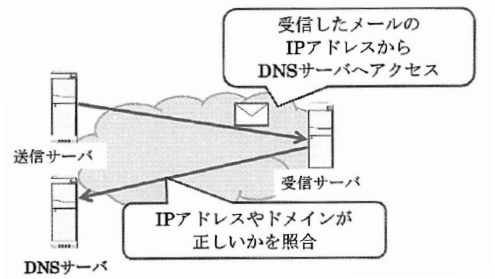

図 3.2　SPF 送信者認証の流れ

表 3.1　SPF 送信者認証の返答

返答	概要
None	ドメインに対する記録なし
Neutral	DNS サーバ管理者による意向で記録なし
Pass	認証成功
Fail	認証失敗
Softfail	認証失敗
Err_temp	一時エラー（クライアントによる不備）
Err_temp	恒久エラー（ドメインなどの不備）

3.2.3　メール作成国と送信元国の不一致

本項目では，メールに表記されている時差の国と送信元の国が一致しない時に時差表記と送信元の国を組み合わせて学習する．これにより，メール作成国とメール送信国が一致しないというスパムメールの特徴を判定に用いる．

一般的なメール送信方法では，メールを作成した国のメールサーバからメールが送信されるので，メールを作成した国と送信した国は一致する．しかしスパムメール送信者が利用する踏み台となるサーバは世界各地に存在している．そのため作成した国から，遠隔操作により他国の踏み台となるサーバを中継して送信することが多い．よって踏み台サーバを中継するメールはメールの作成国と送信国が一致しない場合が多い．そこでこの特徴をベイジアンフィルタの判定に利用する．

メールの作成国の情報は作成された国で作られる Date ヘッダの時差の表記から，その時差に対応する国を得る．また送信国情報は，送信元 IP アドレスより，IP アドレス情報を管理するレジストリの Whois[5]を用いて得る．

メール作成国とメール送信国の不一致に関する概要図を図 3.3 に示す．

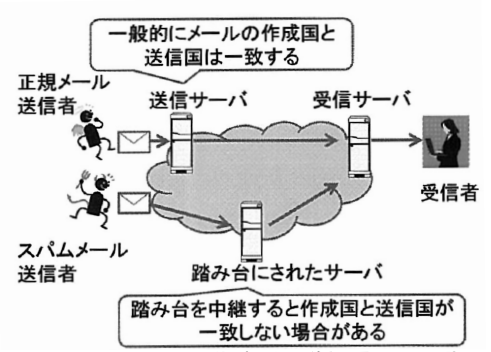

図 3.3　メール作成国・送信国の不一致

3.3　提案方式の処理の流れ

今回提案する方法の処理の流れについて述べる．メール学習時の流れは次の通りである．

① 検査機能にて本システムで追加する項目の検査をおこない，文字列化機能にて検査の結果に対応する文字列に変換する．
② メールヘッダの値として文字列を検査したメールに付加して，メールの学習をおこなう．

この学習方法により，利用者が意識することなく検査項目で得た特徴を学習させることが可能になる．メール判定時の流れは次のとおりである．

① 通常のベイジアンフィルタの判定をおこない，スパムメールまたは正規メールと判定されたメールは確定する．
② ①でどちらにも判定されなかったメールに対して，追加した項目の検査をおこない，文字列を作成する．
③ 学習データベースを参照し，追加した項目のスパムらしさの値を得る．複数の項目がある場合は相乗平均したものをスパムらしさとする．
④ 検査項目より得たスパムらしさの値で再び判定をおこなう．

この検査方法によってベイジアンフィルタの単語による判定ができないメールを，追加項目の判定により補完することが可能になる．

提案するシステムの流れの概要図を図 3.4 に示す.

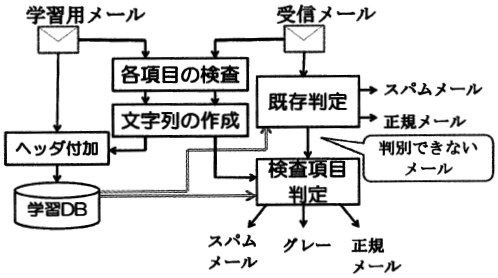

図 3.4　提案手法の処理の流れ

4. 実装方法

4.1 提案方式の実装

提案方式の実装をおこなう. ベイジアンフィルタによる学習・判定の機能は bsfileter[6]を利用する.

送信元情報の検査結果を示す文字列は検査結果それぞれに対し作成される. またベイジアンフィルタでは長い単語を学習する際に, 文字数を短縮する処理がおこなわれる. その短縮の処理方法によっては異なる文字列が同じものとして扱われることがあるため, 本方式で作成する文字列はできる限り短くする.

4.2 検査項目の実装

本節では各検査項目の実装方法と, その結果を文字列に変換する方法について述べる. メールから必要な情報を集め, 検査をおこない文字列を作成する過程を図 3.3 に示す.

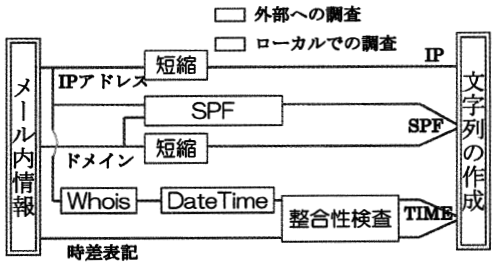

図 3.3　文字列を作成するまでの流れ

4.2.1 送信元サーバの情報の取得

送信元が踏み台サーバであるというスパムメールの特徴を学習・判定に利用するため, 送信元の IP アドレスを表す文字列を作成する. IP アドレスは比較的信頼できるインターネットサービスプロバイダなどのメールサーバが記録した Received ヘッダを参照する. Received ヘッダの from という項目からメールの送信元 IP アドレスを抜き出す. Received ヘッダの例を図 4.1 に示す.

```
Received: from sent_domain [xxx.xxx.xxx.xxx])
        by trasted.ne.jp (Postfix) with ESMTP id xxx
```

図 4.1　標準的な Received ヘッダの例

また文字列に変換する際には, 文字数を短くするために IP アドレスの各オクテッドを 16 進数に変換する. 送信元 IP アドレスを文字列に変換する例を示す.

(IP) 192.168.0.1 → (文字列) IPC0A80001

4.2.2 送信元ドメイン偽装の有無の情報の取得

送信元ドメインの偽装の有無によるスパムメールの特徴を学習・判定に利用するため, 送信元ドメインと送信者認証の結果を表す文字列を作成する. SPF 送信者認証の処理は Perl を用いて実装した.

ここで利用する IP アドレスとドメインは 4.2.1 節と同様に信頼できるメールサーバが記録した Received ヘッダから得る. Received ヘッダの from に記載されているドメインと SPF 送信者認証の結果を組み合わせて文字列に変換する.

また送信元ドメインの文字列を短くするため, ハッシュダイジェストを求め, その最後の 5 文字を抜き出したものをドメインを識別する文字列とする. 以下に文字列を作成する手順の例を示す.

(1) (ドメイン)chuo-u.ac.jp, (IP)192.168.0.1
→ [SPF] → fail

(2) (ドメイン)chuo-u.ac.jp
→ [短縮] → 15c80

(3) fail, 15c80 → (文字列)SPF15c80fail

4.2.3 メール作成国と送信国の不一致の情報の取得

メールの作成国の情報と, 送信国の情報が一致していないという特徴をスパムメール判定に利用するため, メールの作成国の情報と送信国を表す文字列を作成する. メール作成国は Date ヘッダの時差表記から得る. Date ヘッダの例を図 4.2 に示す.

```
Date:  Thu, 24 Jan 2008 11:08:47 +0900
```

図 4.2　Date ヘッダの例

送信国を得るための IP アドレスは 4.2.1 節と同様に信頼できるメールサーバが記録した Received ヘッダより抜き出す.

この 2 つの情報より作成国と送信国の整合性を確認する. まず IP アドレスを管理するレジストリを参照する whois コマンドを用いて, IP アドレスを所有する国を得る. 次に, その国が位置する時差を得て, その時差が Date ヘッダに記載される時差と一致しているか確認する. 一致しなかった場合は時差表記と送信国より文字列を作成する.

以上の作業から Date ヘッダの時差表記と送信国の情報を組み合わせて文字列にする.

文字列を作成する手順の例を示す.

(1) (ip) 192.168.0.1 → [whois] → JP

(2) JP → [DateTime] → +0900
+0900, (時差表記)+0000 → 不一致

(3) JP, +0900
→ (文字列) TIMEJPp0900

5. 評価

5.1 評価方法の概要

評価では実装した提案方式に対し，実際に筆者が受信したスパムメールと正規メールを用いて検出率と誤検出率を求め，ベイジアンフィルタの既存方式による判定結果と比較する．ここで既存方式とは，従来のベイジアンフィルタの判定方法のことを示し，評価ではデフォルト設定の bsfilter を用いて学習・判定した結果と比較する．

まず検査項目を個別でみたときの評価をおこなう．個別の評価はそれぞれの項目の検出率・誤検出率を確認する．また，判定をおこなったメールのうち判定ができたメールの割合を表す判定効率についても確認する．さらに提案方式であるすべての項目を利用したときの検出率・誤検出率の評価をおこなう．

今回の評価ではスパムメールと判定される閾値を 0.9 以上，正規メールと判定される閾値を 0.1 以下とした．また，今回評価に利用したスパムメールと正規メールの数を表に示す．なお学習用メールはそれぞれ受信時間が早かったメール 500 通ずつを利用している．

表 5.1 評価で利用したメール

	学習用	評価用
スパムメール	500	7640
正規メール	500	4623
合計	1000	12263

5.2 評価結果

5.2.1 判定項目個別の評価

本小節では検査項目個別の判定精度を評価する．既存方式による判定結果（既存），IP アドレスの学習による判定結果（IP），送信者認証結果とドメインの学習による判定結果（SPF），メール作成国と送信国の不一致の学習による判定結果（TIME）それぞれにおいて，検出したメールの数，誤検出したメールの数，判定できなかったグレーのメールの数を表 5.2 に示す．また各検査項目の検出率および誤検出率の割合を図 5.1 に示す．さらに，各検査項目の学習がどの程度判定に反映されているかを確認するため，判定することができたメールの割合を図 5.2 に示す．

表 5.2 検査項目個別での評価結果

		既存	IP	SPF	TIME
スパムメール	検出	4041	599	3410	1442
	グレー	3468	0	0	0
	誤検出	131	7	1	75
正規メール	検出	4009	1835	1858	87
	グレー	94	0	0	0
	誤検出	520	16	30	3

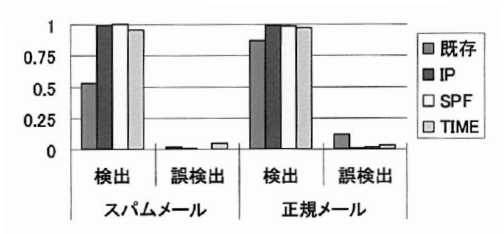

図 5.1 検査項目個別の検出率・誤検出率の割合

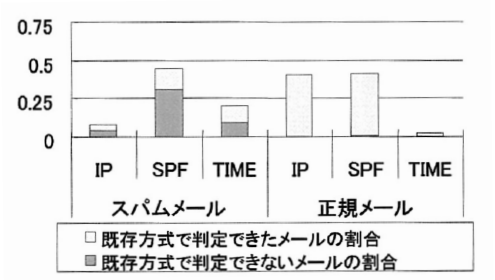

図 5.2 検査項目が判定できたメールの割合

検査項目単体での検出率は，すべての検査項目において，スパムメール，正規メールともに 95%を超えた．一方の誤検出率は IP，SPF の項目では 1%前後だったのに対し，TIME の項目ではやや高く，スパムメールで 4.9%，正規メールで 3.3%だった．

また，検査項目が判定できたスパムメールの割合は，IP, SPF, TIME の項目の順に，7.9%，44.6%，19.9%であった．またそのうち既存方式で判定できなかったメールの割合は同順で 4.3%, 31.0%, 9.6%であった．IP の項目の判定できるメールの割合が低い原因として，他の項目に比べ多パターンが多いことがあげられる．

また正規メールにおいては，既存方式で判定できなかったメールに対しての効果はほとんどなかった．この原因は，既存方式の判定によりすでに 97%以上の正規メールが判定されているためである．

5.2.2 提案方式による評価

本小節では提案方式による判定精度を評価する．評価では，全ての検査項目の組み合わせによる提案方式の精度を，既存方式の精度と比較する．なお，複数の項目を組み合わせる際，スパムメールらしさの値は各項目の値の相乗平均を計算する．既存方式の評価結果（既存）と提案方式の評価結果（提案）について，検出したメールの数，誤検出したメールの数，判定できなかったグレーのメールの数を表 5.3 に示す．さらに検出率および誤検出率の割合を図 5.3 に示す．

表 5.3　提案方式による評価結果

		既存	提案
スパム	検出	4041	6722
メール	グレー	3468	771
	誤検出	131	147
正規	検出	4009	4019
メール	グレー	94	60
	誤検出	520	544

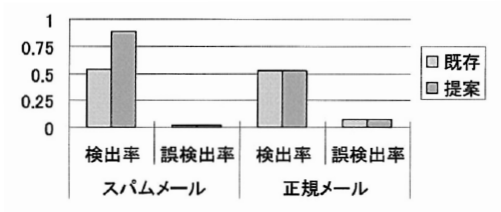

図 5.3　提案方式による検出率・誤検出率の割

提案方式によって検出されたスパムメールは既存方式から 35.1%増加したのに対し，誤検出率は 0.2%の増加だった．また，提案方式によって検出された正規メールは，既存方式から 0.2%の増加，誤検出率は 0.5%の増加となった．

また，既存方式で判定できなかったスパムメールの 77.3%を検出した．提案方式で追加した情報により，ベイジアンフィルタの補完的な判定ができているといえる．

6. おわりに

本稿では，ベイジアンフィルタの精度向上を目的として，送信元に関する三通りの検査により得た情報をベイジアンフィルタのスパムメール判定に利用する方法を提案した．実際のメールを用いた評価をおこなった結果，既存のベイジアンフィルタの判定結果と比較してスパムメールの検出数の割合を 35%上昇することができた．一方，誤検出率の上昇は 1%未満であった．またベイジアンフィルタで判定できなかったスパムメールを提案方式で 77%検出することができた．以上より，本提案方式の一定の有効性を示せた．今後の課題としては次をあげる．

- 新たなパラメータを追加し，スパムメールの検出率を向上させ，誤検知率の低減させる．今回は送信元の特徴を重視したが，他の特徴と組み合わせることで補完効果がより高くなると考えられる．
- 検査項目において，学習したメールが判定に利用される割合である判定効率が上がるようにする．例えば IP の項目における検出率は高いが，判定の対象となるメールが少なかった．学習する範囲をネットワークアドレスに拡大するなどの処理により判定効率の向上が期待できる．

参考文献

[1]MessageLabs Intelligence: 2007 Annual SecurityReport
http://www.messagelabs.com/mlireport/MLI_2007_Annual_Security_Report.pdf
[2] A Plan for Spam
http://www.paulgraham.com/spam.html
[3] Simple Mail Transfer Protocol
RFC2821RFC821，974，1869
http://www.ietf.org/rfc/rfc2821.txt
[4] Sender Policy Framework
RFC4408 April 2006
http://www.ietf.org/rfc/rfc4408.txt
[5] 日本レジストリサービス，http://jprs.jp
[6] bsfilter，http://bsfilter.org/