情報処理学会論文誌　Vol. 49　No. 9　3093–3103 (Sep. 2008)

ベイジアンフィルタにおける画像スパムのフィルタリング方式の実現

上村昌裕[†1]　田端利宏[†1]

インターネットの普及とともに，迷惑メールの増加が近年問題となっている．2006年には，迷惑メールが電子メール全体の91%を占めたとの調査結果も存在する．迷惑メール対策として，ベイズ理論を用いて統計的にフィルタリングを行うベイジアンフィルタが広く利用されている．その特徴として，フィルタリングの精度が高く，迷惑メールの流行や個人の嗜好に合わせたフィルタリングが行えることがある．しかし，その回避策として，迷惑メールの内容を画像化して送信する画像スパムが急増している．ベイジアンフィルタはテキストデータに対して学習と判定を行うので，画像などのバイナリデータに対しては，適切な学習と判定ができない．そこで，本論文では，画像スパム対策として，ファイルサイズなどの添付画像の情報に着目し，これらの情報を既存のベイジアンフィルタのコーパス（学習データ）に加え，フィルタリングを行う方式を提案する．また，その評価結果を報告する．

A Bayesian-filter-based Image Spam Filtering Method

Masahiro Uemura[†1] and Toshihiro Tabata[†1]

In recent years, with the spread of the Internet, the increase in the number of spam has become one of the most serious problems. A recent report reveals that 91% of all e-mail exchanged in 2006 was spam. Using the Bayesian filter is a popular approach to distinguish between spam and legitimate e-mails. It applies the Bayes theory to identify spam. This filter proffers high filtering precision and is capable of detecting spam as per personal preferences. However, the number of image spam, which contains the spam message as an image, has been increasing rapidly. The Bayesian filter is not capable of distinguishing between image spam and legitimate e-mails since it learns from and examines only text data. Therefore, in this study, we propose an anti-image spam technique that uses image information such as file size. This technique can be easily implemented on the existing Bayesian filter. In addition, we report the results of the evaluations of this technique.

1. はじめに

迷惑メールは，近年のインターネットの普及とともに，大きな社会問題となっている．送信に要する費用の少なさから，その数は年々増加している．2001年には，電子メール全体の10%以下だった迷惑メールが，2003年には50%を上回り，2004年には65%，2006年には91%を占めたとの調査結果が存在する[1)]．迷惑メールの増加による問題点として，正当な電子メールと迷惑メールの仕分けにかかる時間，迷惑メールによる記憶領域の使用，および通信回線を流れる転送データ量の増加による電子メールの通信遅延があげられる．また，最近では，フィッシングメールと呼ばれる詐欺メールも急増している[2)]．これらの問題から，電子メールの信頼性や利便性が低下している．このため，電子メールの信頼性を保つうえで，迷惑メールを排除するための技術的対策が必要となっている．

迷惑メールに対する技術的対策の1つとして，ベイジアンフィルタがある．ベイジアンフィルタは，過去に受信したメールから，統計的に単語（トークン）の迷惑メール確率を計算して学習する．このようにして作成した学習データ（コーパス）をもとに，新しく受信した電子メールが，正当な電子メールであるか迷惑メールであるかを推定する方式である．ベイジアンフィルタは，フィルタリング精度が高く，近年利用が増えている．

しかし，ベイジアンフィルタの回避策として，迷惑メールの内容を画像化して送信する電子メール（以降，画像スパムと略す）が急増している．McAfeeによると，2006年末には，画像スパムは迷惑メール全体のうちの65%を占めている[3)]．ベイジアンフィルタはテキストデータに対して学習と判定を行うため，画像のようなバイナリデータに対しては，学習と判定ができない．このため，画像スパムは，テキストスパムに比べ，フィルタを回避する場合が多い．具体的な画像スパムの例として，正当な電子メールであるかのような内容を本文に載せ，迷惑メールの内容を画像として送信することで，フィルタを回避するものがある．

そこで，本論文では，画像スパム対策として，ファイルサイズなどの添付画像の情報に着目し，これらの情報を既存のベイジアンフィルタのコーパスに加え，フィルタリングを行う方式を提案する．提案方式の利点は，テキストデータのみを学習と判定の対象としているベイジアンフィルタの実装方式（以降，従来方式と呼ぶ）と比較して，誤検出（正当な電子メールを誤って迷惑メールと見なすこと）を増やすことなく，画像スパムの見逃し（迷惑

†1 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University

メールを誤って正当な電子メールと見なすこと）を減らすことができる点である．また，提案方式の処理時間は，従来方式と比べて，問題となるほど増加しない．

提案方式は，画像が添付された電子メールを判定するとき，最初に従来方式で判定する．この判定結果が正当な電子メールであれば，正当な電子メールとして判定を確定させる．これは，正当な電子メールを十分に学習させたコーパスを用いて正当な電子メールを判定させる場合，その電子メールのヘッダや本文に現れる特徴が，コーパス中の正当な電子メールの特徴と一致する場合が多いため，ほとんどの場合，スパム確率が十分に低く，ヘッダと本文の評価で正当な電子メールと判定できるからである．一方，従来方式での判定結果が，迷惑メールと疑われる場合，さらに画像情報を加えて判定する．このように，疑わしいメールについてのみ，添付画像の特徴を加えて評価するため，画像付きの正当な電子メールの誤検出を防止でき，判定精度を向上させることができる．

本論文の構成は以下のとおりである．2 章でベイジアンフィルタについて述べ，3 章で画像スパムの現状と問題点について述べる．次に，4 章で画像スパムに添付されている画像の調査結果と提案方式について述べる．5 章では，提案方式の実装内容と評価について述べ，6 章で関連研究について述べる．最後に 7 章で本論文のまとめを述べる．

2. ベイジアンフィルタ

2.1 概　　要

ベイジアンフィルタは，過去に受信した迷惑メールと正当な電子メールのデータを基に，新たに受信した電子メールが迷惑メールであるか正当な電子メールであるかを推測する手法である．新たな電子メールを受信すると，テキストをトークン単位に分割し，コーパスを基に，各トークンの迷惑メール確率を計算する．次に，これらの確率を基に，判定対象の電子メールの迷惑メール確率を計算する．それから，この確率があらかじめ設定した閾値を上回った場合に迷惑メールと判定し，下回った場合に正当な電子メールと判定する．迷惑メール確率の計算方法として，Graham 方式[4] や Robinson 方式[5),6)] が多く用いられている．これらの方式において，トークンと電子メールに対する迷惑メール確率は，0 から 1 の間の値をとる．確率が 0 に近い値は，そのトークンや電子メールが，正当である可能性が高いことを意味する．確率が 1 に近い値は，迷惑メールである可能性が高いことを意味する．

2.2 Robinson 方式

提案方式では，Robinson 方式を用いた．Robinson 方式では，トークンごとの迷惑メール確率 $f(w)$ を以下のように求める．

$$p(w) = \frac{\frac{b}{n_{bad}}}{\frac{g}{n_{good}} + \frac{b}{n_{bad}}} \tag{1}$$

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

- g：正当な電子メールにおけるトークン w の出現回数
- b：迷惑メールにおけるトークン w の出現回数
- n_{good}：正当な電子メール数
- n_{bad}：迷惑メール数

x は今まで 1 度もメール中に出現していないトークンが，迷惑メールで最初に出現する予測確率とし，s をその予測に与える強さとする．また，n はトークン w が出現したメール数とする．x と s の値は，パフォーマンスを最適化するためのテストにより，$x = 0.5$，$s = 1$ が妥当であるとされている．

Graham 方式と Robinson 方式の主な違いは，トークン w の出現回数が少ない場合の対処法である．Graham 方式では，トークン w が迷惑メールのみに数回出現した場合，そのトークンの迷惑メール確率 $p(w)$ が 1 になる計算方法となっている．この場合，そのトークン w に最大の迷惑メール確率を与えるには情報が少ないといえる．

Robinson 方式はこの問題を解決するため，トークン w の出現回数が少ない場合，$p(w)$ の比重が小さくなる計算方法をとり，トークン w の情報が十分でないことを $f(w)$ に加えることができる．学習数が増えるにつれ，出現回数 n が大きくなっていき，$f(w)$ の値は漸近的に $p(w)$ の値に近づいていく．また，トークン w の出現回数が 0 の場合，そのトークンの迷惑メール確率は 0.5 となる．さらに，判定対象のメールが迷惑メールである確率は次の I で与えられる．

$$H = C^{-1}\left(-2ln\prod_{w}(1 - f(w), 2n)\right) \tag{3}$$

$$S = C^{-1}\left(-2ln\prod_{w}(f(w), 2n)\right) \tag{4}$$

$$I = \frac{1 + H - S}{2} \tag{5}$$

C^{-1} は逆 χ^2 関数（inverse chi-squpre 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（Indeicator）である．

2.3 特　　徴

ベイジアンフィルタの主な特徴は，以下の 3 点がある．これらの特徴から，ベイジアンフィルタは，多くの迷惑メールを検出することができ，利用が増えている．

(1)　精度向上
　学習させる電子メール数が増加すればするほど，フィルタリング精度が向上する．

(2)　柔軟性
　受信する電子メールや迷惑メールは，各利用者ごとに異なり，また，同じ利用者でも時期によって異なる．ベイジアンフィルタは，学習させることで，内容の変化とともにフィルタリング基準も変化し，各利用者・各時期の傾向に合ったフィルタリングを行うことができる．

(3)　利便性
　学習データを基に判定するので，キーワード指定といった作業をしなくてよい．

しかし，迷惑メールの送信者側もベイジアンフィルタを通過できるように，迷惑メールの内容を工夫するようになっている[7]．なかでも，近年，画像スパムが急増し，問題となっている．

3. 画像スパム

3.1 画像スパムの現状

McAfee によると，画像スパムは 2005 年ごろから登場し，増加し続けている[3]．2006 年の初めには，迷惑メール全体のうち画像スパムが占める比率は 30%であったが，10 月には 40%，2006 年末には 65%とその数は増加し続けている．ベイジアンフィルタに代表されるテキストベースのフィルタリングは，精度が高いものの，画像に書かれている単語や文字列，文章は抽出することができず，画像に対しての対策がとれない．したがって，既存のベイジアンフィルタでは，画像スパムに対してヘッダと本文のみで判定している．

迷惑メール送信者は送信したい内容を画像化しているため，本文は何も書かない場合や短い場合が多く，ヘッダの情報が判定に大きく影響する．しかし，ヘッダは SMTP の設計上，改変を行うことができ，確実な信頼性があるとはいえない．また，本文が含まれる場合でも，Word Salad[8),9)] のように，迷惑メールの内容とまったく関係のない内容，あるいは正当な電子メールであるかのような内容で送信する場合が多い．したがって，画像スパムはテキストベースの迷惑メールに比べ，フィルタリングを通過する場合が多い．また，画像スパムのサイズは，テキストベースの迷惑メールに比べ，約 3〜4 倍大きいため，メールサーバへの負担が大きいのも問題である[3]．

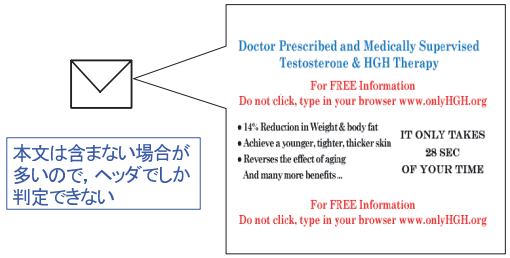

図 1　画像スパムの添付画像の例
Fig. 1　Example of an attached image contained in an image spam e-mail.

画像スパムの手法は，テキストのみの迷惑メールと同様に次々変化している．主に，OCR（光学式文字読取装置）による検出を回避する手法と，シグネチャによる検出を回避する手法の 2 つのタイプに分類される[10]．前者は，ユーザが人間であるかどうかを認証する CAPTCHA[11] 技術を利用して，画像をゆがませるなどの加工を加え，人間の目には読み取れるが，OCR では読み取れない画像を送信する手法である．後者は，画像の背景に加えられたノイズやファイル名，サブジェクト名をランダムに変化させ，人間の目には同じように見える画像でも，ランダム化された画像を用いることで，シグネチャによる検出を困難にする手法である．また，アニメーション GIF やマルチレイヤのイメージファイルを用いて，フィルタから宣伝メッセージを隠そうとする手段もある．

画像スパムの例を **図 1** に示す．画像スパムは図 1 のように内容を画像化し，本文は何も書かない，あるいは正当な電子メールであるかのような文章を書いている場合が多い．

3.2 画像スパムに添付される画像の調査

2006 年 5 月から 2007 年 2 月までの期間に，研究室の構成員の 1 人が受信した迷惑メール 10,131 通を調査した．調査結果を**表 1** に示す．迷惑メールのうち画像スパムは 2,250 通（22.2%）あり，添付されていた画像数の合計は 2,429 個であった．添付画像の画像形式は，GIF，JPEG，PNG の 3 種類があり，GIF が全体の 6 割を占めている．また，言語別でみると，本文が英語の画像スパムが全体の 99%を占めている．文献 10) の調査結果では，GIF

表 1　画像形式の内訳
Table 1　Details pertaining image formats.

画像形式	本文が英語	本文が日本語	合計
GIF	1,557	2	1,559
JPEG	814	15	829
PNG	41	0	41
合計	2,412	17	2,429

が全体の 85%を占め，英語の画像スパムが 91%を占めている．これらの調査結果より，画像スパムに添付される画像は，英語の GIF 画像スパムが多いといえる．現在は，英語の画像スパムが多いが，今後は日本語の画像スパムも増加することが考えられる．

4. 提案方式

4.1 概　　要

提案方式は，既存のベイジアンフィルタに画像情報を学習させ，このコーパスを用いて新たな電子メールを判定させる手法である．これにより，画像スパムに対するフィルタリング精度を上げることを目的としている．

4.2 コーパスに組み込む画像情報

ベイジアンフィルタのフィルタリング精度を上げるうえで，コーパスに組み込む画像情報の検討は非常に重要である．テキストのみの正当な電子メールと比べると，画像が添付された正当な電子メールは，数が少ない傾向にあると考えられる．これに対し，画像スパムは，テキストフィルタリングの回避策として増加している．したがって，画像が添付されている電子メールは画像スパムである可能性が高い．しかし，正当な電子メールに画像を添付して送信する場合もあるので，画像付きの電子メールを画像スパムと判定することは問題がある．

そこで，提案方式では，画像スパムの画像のメタデータをコーパスに学習させ，判定を行うことにより，画像スパムの見逃しを減らし，フィルタリング精度を上げる．また，テキストフィルタリングと連携させることで，誤検出を減らすことも実現する．

画像スパムに添付される画像は，テキストとして載せる情報を画像化するので，文字を多く含む場合が多い．しかし，画像は一般的に，絵や写真，イラストなどテキストでは表現できない情報を画像で表現する場合が多い．したがって，画像のメタデータに違いがあると考えられる．表 1 から，画像スパムは，GIF や JPEG の画像形式を利用する場合が多いことが分かる．GIF や JPEG は，画像のファイルサイズを小さくするために画像を圧縮するので，色素が少ない画像や単純な画像ほど圧縮率が高い．したがって，文字が多い画像は，絵やイラストなどに比べ，圧縮率が高いと考えられる．提案方式は，画像の内容を解析せずに利用できる情報を用いることとし，コーパスに追加するメタデータをファイル名，ファイルサイズ，面積，圧縮率の 4 つとした，各情報に現れる特徴について述べる．

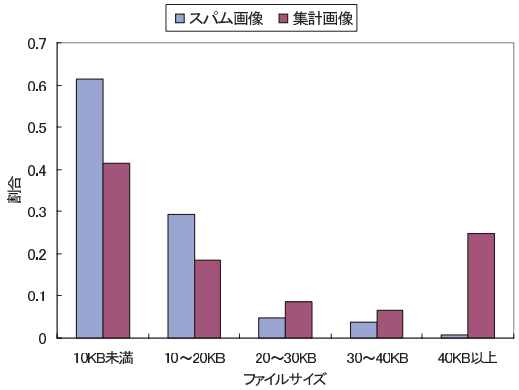

図 2　スパム画像（GIF）と集計画像のファイルサイズ
Fig. 2　File sizes of spam (GIF) and conventional images.

4.3 GIF 画像の分析

分析に用いた画像は，表 1 の GIF 画像（1,559 個）と，それらの面積の分布を基準に，インターネット上に存在する画像を検索エンジンの Google [12] で検索して集計した GIF 画像（784 個）である．各情報の調査結果をファイルサイズは**図 2**，面積は**図 3**，圧縮率は**図 4** に示し，各情報に現れる特徴について述べる．以降，画像スパムに添付されている画像をスパム画像，集計した GIF 画像を集計画像と略す．

4.3.1 ファイル名

正当な利用者が，電子メールに画像を添付して送信する際，同じ画像を同じ受信者に何度も送信することはあまり考えられない．しかし，迷惑メール送信者は，迷惑メールを大量に何度も送信するので，同じ画像を何度も送信することになる．したがって，画像のファイル名を学習させることで，同じ画像のファイル名が送られてきた場合，送られてきた画像スパムの迷惑メール確率を上げることができる．実際，調査した画像スパムの中で，同じファイル名の画像を送信するものもいくらかあった．

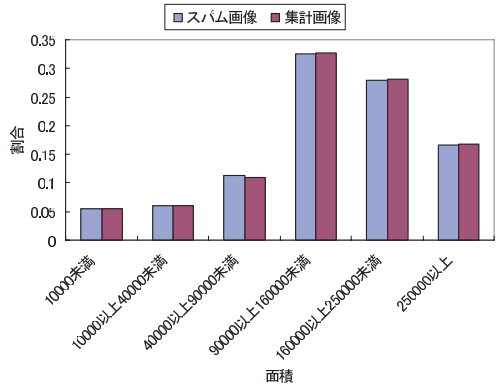

図 3　スパム画像（GIF）と集計画像の面積
Fig. 3　Areas of spam (GIF) and conventional images.

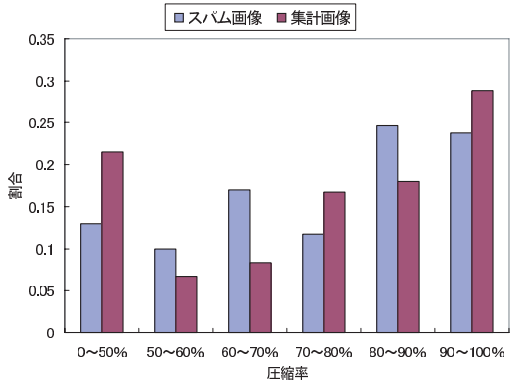

図 4　スパム画像（GIF）と集計画像の圧縮率
Fig. 4　Compressibility of spam (GIF) and conventional images.

4.3.2　ファイルサイズ

図 2 より，画像スパムの添付画像は，ファイルサイズが小さいものが多い．面積の分布を揃えて比較しているため，この結果は，画像スパムの添付画像の圧縮率が高く，より単純な画像であることがいえる．20 KB 以下の画像が，全体の 9 割以上を占めているのは着目すべき点であり，集計画像と大きな違いがある．

4.3.3　面　　積

画像を認識するためには，人間の目に見えやすいようにある程度の大きさが必要である．これは，画像スパムの画像でも，絵やイラストなどの画像でも同じことがいえる．したがって，面積の分布で大きな違いがあるとは考えられないので，図 3 のように分布を揃え，他の情報を比較できるようにした．

4.3.4　圧　縮　率

迷惑メールの内容を画像にする場合，画像の面積はある程度の大きさが必要である．また，画像スパムのファイルサイズは小さい傾向にある．図 4 から，スパム画像と集計画像の圧縮率の分布に差があることが分かる．圧縮率が 50%以上の画像の場合，スパム画像はスパム画像全体の 87%，集計画像は集計画像全体の 78%を占めており，スパム画像の方が圧縮率が高いファイルの割合が多い．また，文献 13) の調査では，圧縮率が 50%以上のスパム画像は，スパム画像全体の約 93%を占め，スパムでない画像では，スパムでない画像全体の約 42%を占めており，圧縮率による分布に大きな差があることが報告されている．したがって，画像スパムの添付画像は集計画像より圧縮率が高いといえる．

4.4　JPEG 画像の分析

分析に用いた画像は，表 1 の JPEG 画像（829 個）と，携帯電話のカメラで撮影した画像（以降，携帯画像）とした．携帯画像を分析した理由は，正当な送信者が画像を送信する際に，最も利用する機会が多いと考えられるからである．総務省の調査によると，2006 年 12 月末の携帯電話の加入契約数は約 9,500 万台であり[14)]，最近の携帯電話は，ほぼすべての機種にカメラが内蔵されている．したがって，携帯で画像を撮影する機会は多く，それにともない，画像を送信する機会も多いと考えられる．著者の携帯電話のカメラで撮影した JPEG 画像の詳細を**表 2** に示し，画像スパムに添付されていた JPEG 画像の詳細を，ファイルサイズ，面積，圧縮率ごとにそれぞれ**表 3**，**表 4**，**表 5** に示す．

これらの表から，スパム画像は，ファイルサイズが 10 KB 以上 20 KB 未満で，面積が 240 × 320 以上 480 × 640 未満で，圧縮率が 90%以上の画像が多い．携帯画像で面積が 240 × 320 以上 480 × 640 未満の場合，ファイルサイズは 30 KB 以上で圧縮率は 90%前後である．したがって，スパム画像は，携帯画像よりファイルサイズが小さく，圧縮率が高い傾向があることが分かる．前節での GIF のスパム画像と集計画像の比較では，面積の分布を揃えた場合，スパム画像の方がファイルサイズが小さく，圧縮率が高い傾向があることを示した．これらより，JPEG のスパム画像と携帯画像を比較したときの傾向と GIF のスパム画像と集計画像を比較したときの傾向は類似しているといえる．

表 2　携帯画像（JPEG）の詳細
Table 2　Dedails mobile images (JPEG).

面積	ファイルサイズ（KB）	圧縮率（%）
120 × 160	11	81
240 × 320	34	85
480 × 640	113	92
768 × 1, 024	218	92
960 × 1, 280	294	91
1, 200 × 1, 600	402	92

表 3　スパム画像のファイルサイズ（JPEG）
Table 3　File sizes of spam (JPEG).

ファイルサイズ	10 KB 未満	10〜20 KB	20〜30 KB	30 KB 以上
数	35	777	16	1

表 4　スパム画像の面積（JPEG）
Table 4　Areas of spam (JPEG).

面積	120 × 160 未満	120 × 160 〜240 × 320	240 × 320 〜480 × 640	480 × 640 以上
数	0	30	797	2

表 5　スパム画像の圧縮率（JPEG）
Table 5　Compressibility of spam (JPEG).

圧縮率	80%未満	80〜90%	90%以上
数	0	5	824

4.5　画像情報のトークン

ファイルサイズ，面積，圧縮率は，数値として得る情報である．数値としてコーパスに学習させた場合，まったく同じ値でないと同じトークンとして学習されない．また，判定時もまったく同じ値でないと学習された情報を利用できない．したがって，数値として学習や判定に用いるより，ある程度の範囲でまとめて 1 つのトークンとし，学習と判定に用いる方が効果的に迷惑メール判定確率に反映できると考えられる．画像情報のコーパス組み込み例として，ファイルサイズが 10 KB〜20 KB の場合，“size10_20 KB” といったトークンとしてコーパスに組み込む．以降，この画像情報のトークンを画像トークンと呼ぶ．

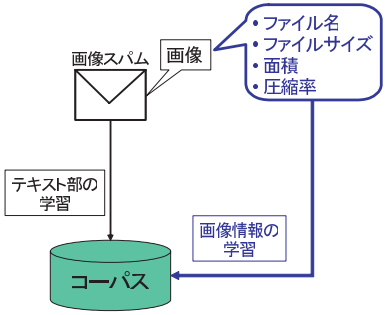

図 5　学習時の流れ
Fig. 5　Learning flow.

4.6　処理の流れ

4.6.1　学　習　時

学習時の流れを**図 5** 示す．従来方式は，画像スパムのヘッダと本文のみをコーパスに学習させていた．提案方式では，これらに加えて，ファイル名などの画像情報をコーパスに学習させる．

4.6.2　判　定　時

これまでの調査により，画像が添付されたメールは，以下の 2 つの傾向がみられることが多い．

(1)　画像が添付された正当な電子メールは，テキストのみで判定すると，その迷惑メール確率は，画像が添付されていない正当な電子メールと同じぐらい低い．これは，画像が添付されていない正当な電子メールと同等の本文が記述されているためである．

(2)　画像スパムは，テキストのみで判定すると，その迷惑メール確率は，低くても 0.5 程度である．トークンの学習数が少ない場合，そのトークンの迷惑メール確率は 0.5 に近い値となり，それらのトークンを多く含む電子メールの迷惑メール確率も 0.5 に近い値となる．画像スパムの場合，送信元は様々であるため，ヘッダは学習数が少ないトークンとなる．本文を含まない場合はヘッダにより判定されるため，その画像スパムの迷惑メール確率は 0.5 に近くなる．本文が Word Salad を利用した内容であった場合，正当であると学習させたトークンが本文中に多く含まれていれば，その画像スパムの迷惑メール確率は 0.5 より小さくなる．反対に，Word Salad 中に正当であると学習されたトークンが少なければ，その画像スパムの迷惑メール確率は 0.5 に近い

値となる．我々の調査では，後者の傾向がみられた．

これらの傾向を基に，画像が添付されたメールを判定する方式の流れを**図 6** に示し，以下でその手順を述べる．

(1)　テキストのみで判定する従来方式で迷惑メール確率を計算する．

(2)　迷惑メール確率が s 未満の場合，正当な電子メールと判定する．s 以上 t 未満の場合，画像情報も加えて判定する提案方式で再度迷惑メール確率を計算する．テキストのみの判定確率が t 以上の場合，画像スパムでも，そのメールは画像の有無に関係なくほぼ迷惑メールであると考えられるため，再計算は行わなず，迷惑メールと判定する．

(3)　提案方式で計算した迷惑メール確率が閾値 t 以上の場合，迷惑メールと判定し，t 未満の場合，正当な電子メールと判定する．

提案方式の評価では，閾値 s を 0.4 と設定した．これは，我々の調査では，画像スパムの迷惑メール確率は，低くても 0.5 程度であったこと，正当な電子メールの迷惑メール確率は，すべて 0.4 未満であったことを考慮したうえでのパラメータ設定である．また，正当な電子メールに多く含まれるトークンと迷惑メールに多く含まれるトークンが同程度含まれる場合，迷惑メール確率は 0.5 前後に分布することや，学習されていないトークンの迷惑メール確率の初期値が 0.5 であることも考慮し，0.5 よりも低く設定している．閾値 t については，ベイジアンフィルタは，経験的にパラメータを設定することが多いため，ベイジアンフィルタのデフォルトの設定でよく使用される 0.9 と設定した．これらのことから，画像が添付されたメールの迷惑メール確率が 0.4 以上 0.9 未満の場合，確率を再計算する．

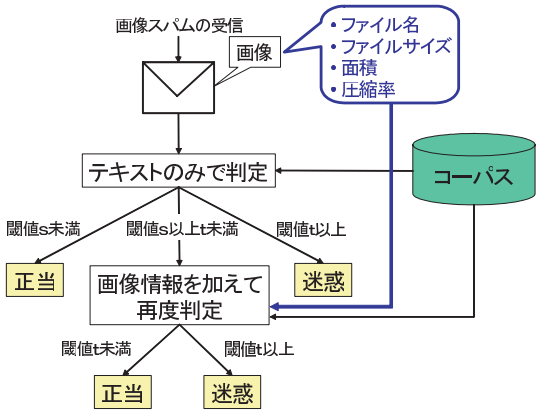

図 6　判定時の流れ
Fig. 6　Judgement flow.

5. 実装と評価

5.1 実 装 内 容

提案方式は，ベイジアンフィルタを採用している bsfilter [15] に実装した．添付画像の画像形式は，GIF，JPEG および PNG の画像形式に対応している．オリジナルの bsfilter は，テキストデータのみを学習と判定の対象としているため，以降では，従来方式と表記する．

5.2 実験に用いた電子メール

実験に用いた電子メールは，研究室の構成員の 1 人が受信した英語の正当な電子メールと，英語の迷惑メールおよび英語の画像スパムである．ここで，迷惑メールとは，画像スパムでない迷惑メールのことを指す．学習時と判定時に用いた電子メールの数を**表 6** に示す．画像スパムは GIF 画像を添付したものである．学習時と判定時に用いた電子メールは異なるものを使用しているので，合計 1,400 通の電子メールを実験に用いた．今回の実験は，正当な電子メールに画像が添付されていない場合について行った．これは，正当な電子メールには画像が添付される場合が少ないことと，画像情報をコーパスに組み込むことによる画像スパムの迷惑メール判定確率の変化を調査するためである．

5.3 コーパスに組み込む画像トークン

各画像情報は，前節での調査を基に，**表 7**，**表 8**，および**表 9** に示す範囲で 1 つのトークンとし，コーパスに組み込んだ．添付されていた画像は 201 個であり，組み込まれたトークン数の内訳を示す．

5.4 確率計算に用いる画像トークンについて

今回の実験は，Robinson 方式で迷惑メール判定確率の計算を行った．Robinson 方式では，メール中に現れるすべてのトークンを迷惑メール判定確率の計算に用いる．今回の実験

表 6　実験に用いた電子メール（本文の言語は英語）
Table 6　E-mails used for the experiment (English-language text).

	学習時	判定時
正当な電子メール	300	300
迷惑メール	200	200
画像スパム	200	200
合計	700	700

表 7　ファイルサイズの画像トークン
Table 7　Image tokens of file sizes.

ファイルサイズ	0〜10 KB	10〜20 KB	20〜30 KB	30〜40 KB	40 KB 以上
トークン名	I_size10 KB	I_size10_20 KB	I_size20_30 KB	I_size30_40 KB	I_size_40 KB
トークン数	98	78	12	5	8

表 8　面積の画像トークン
Table 8　Image tokens of areas.

面積（pixel）	10,000 未満	10,000 〜40,000	40,000 〜90,000	90,000 〜160,000	160,000 〜250,000	250,000 以上
トークン名	I_area100	I_area200	I_area300	I_area400	I_area500	I_areaBig
トークン数	3	6	11	68	38	75

表 9　圧縮率の画像トークン
Table 9　Image tokens of compressibility.

圧縮率	0〜50%	50〜60%	60〜70%	70〜80%	80〜90%	90〜100%
トークン名	I_compress 50	I_compress 50_60	I_compress 60_70	I_compress 70_80	I_compress 80_90	I_compress 90_100
トークン数	27	22	18	23	50	61

では，画像トークンは迷惑メールにしか現れないので，画像トークンの迷惑メール確率は高い．しかし，メール中の全トークンに占める割合が小さければ，いくら画像トークンの迷惑メール確率が高くても，そのメールの判定確率にはあまり反映されない．また，画像スパムの特徴を持つ画像トークンでなければ，判定時に迷惑メール確率を上げることはない．したがって，判定に用いる画像トークンの数を変化させて確率を計算した．

画像トークン数は，数値を設定して組み込むのではなく，全トークン数の割合で組み込むことにより，メールごとに偏りがないようにする．追加した画像トークンの割合による影響を明らかにするため，テキストの全トークン数に対する追加した画像トークン数の割合を 10%，30%，および 50%とした．指定した割合の画像トークンを追加する方法を以降で述べる．画像トークンのうち，ファイル名は数値ではなくテキストデータであり，画像そのものの特徴を表すものではないため，本文のテキストと同様に 1 つのトークンを追加する．ファイル名以外の 3 つの画像トークン（ファイルサイズ，面積，圧縮率）については，追加する画像トークン数の合計が，テキストのみの全トークン数の指定した割合の個数になるように，各画像トークン数を算出する．なお，算出した値の小数点以下は切り捨て，3 つの画像トークンは，同じ割合で追加する．

表 10　提案方式で計算した画像スパムのスパム確率の分布
Table 10　Distributions of spam probability calculated by our proposed method.

判定確率	0.5 以上 〜0.6 未満	0.6 以上 〜0.7 未満	0.7 以上 〜0.8 未満	0.8 以上 〜0.9 未満	0.9 以上 〜1.0 未満	1.0
従来方式	17 (8.5%)	6 (3.0%)	4 (2.0%)	0 (0.0%)	25 (12.5%)	148 (74.0%)
10%追加	10 (5.0%)	1 (0.5%)	4 (2.0%)	4 (2.0%)	33 (16.5%)	148 (74.0%)
30%追加	4 (2.0%)	1 (0.5%)	0 (0.0%)	2 (1.0%)	43 (21.5%)	150 (75.0%)
50%追加	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.5%)	35 (17.5%)	164 (82.0%)

(1)　判定前に，追加する画像トークン数の割合をあらかじめ設定する．

(2)　従来方式で迷惑メール確率を計算するときに，判定に用いたテキストのトークン数を数える．

(3)　提案方式で迷惑メール確率を計算するときに，あらかじめ設定した割合とテキストのトークン数をもとに，追加する画像トークン数を式 (6) で算出する．

$$(\text{各画像トークン数}) = (\text{テキストのみの全トークン数}) \times (\text{割合})/3 \qquad (6)$$

(4)　算出した画像トークン数分の画像トークンを追加し，迷惑メール確率を計算する．

例として，従来方式で判定したときのテキストのトークン数が 100 で，判定に用いる画像トークン数を 10%追加としたときについて述べる．この場合，各画像トークン数は，3 個（$100 \times 0.1/3$）となる．この結果，提案方式の判定時に用いられる全トークン数は 110 個（ファイル名：1 個，ファイルサイズ：3 個，面積：3 個，圧縮率：3 個）となる．

5.5　評　　価

判定結果を**表 10** に示す．閾値を 0.9 としたときの画像スパムの見逃し率は，表の上から順に，13.5%（従来方式），9.5%（10%追加），3.5%（30%追加），0.5%（50%追加）である．画像トークン数の 10%追加では，従来方式と比べ，4%の改善がみられたが，見逃し率は 9.5%と高い．また，50%追加では，見逃し数はかなり改善されるものの，画像情報の影響を受けやすく，誤検出が増加すると考えられる．これは，50%追加では，従来方式で見逃した 27 個の画像スパムの内の 10 個の判定確率を 0.9 以上に，16 個を 1.0 にまで上げているため，画像が添付された正当な電子メールの判定確率が 0.5 付近の場合でも，誤検出になる可能性が高くなるからである．

表 11　使用する画像トークンの種類を変化させたときの判定確率の分布
Table 11　Distributions of the judgment probability by the change of the kind of the image token to use.

判定確率	0.5 以上 ～0.6 未満	0.6 以上 ～0.7 未満	0.7 以上 ～0.8 未満	0.8 以上 ～0.9 未満	0.9 以上 ～1.0 未満	1.0
全種類	4	1	0	2	43	150
	(2.0%)	(0.5%)	(0.0%)	(1.0%)	(21.5%)	(75.0%)
ファイルサイズのみ	3	2	0	2	41	152
	(1.5%)	(1.0%)	(0.0%)	(1.0%)	(20.5%)	(76.0%)
面積のみ	3	2	0	2	41	152
	(1.5%)	(1.0%)	(0.0%)	(1.0%)	(20.5%)	(76.0%)
圧縮率のみ	3	2	0	2	42	151
	(1.5%)	(1.0%)	(0.0%)	(1.0%)	(21.0%)	(75.5%)

提案方式は，判定確率が 0.5 以上閾値未満の画像付きメールの判定確率を，設定した閾値まで上げることが目的であり，1.0 まで上げることではない．30%追加では，従来方式で見逃した 27 個のうち，18 個を 0.9 以上に，2 個を 1.0 にまで上げているので，十分に迷惑メール判定確率を上げているといえる．

また，提案方式は，最初にテキストのみで判定を行う．実験では，テキストのみの判定で，閾値 0.4 を下回るの画像スパムはなかったため，従来方式と誤検出率は変わらなかった．つまり，提案方式は，見逃し率を下げ，かつ誤検出率が従来方式とほぼ同じであるので，従来方式よりも精度が高いといえる．これらの結果より，閾値を 0.8 とすると，さらに見逃し率が低くなることが分かる．よって，閾値を 0.8 とし，全トークン数の 30%の画像トークン数を追加するのがこの場合の適切な値の設定であると考えられる．このとき，見逃し率は 2.5%である．

ファイルサイズ，面積，圧縮率が個別に与える影響を調査するため，各画像トークンのみを全トークン数の 30%追加したときの判定結果を**表 11** に示す．表 11 から，閾値を 0.8 と設定した場合，いずれの場合も見逃し率は 2.5%であることが分かる．このことから，各画像トークンがうまく画像スパムの特徴をとらえていることが分かる．

処理時間に関しては，Pentium III（1.26 GHz）搭載の計算機を用いて測定を行った．結果を**表 12** に示す．提案方式は従来方式に比べ，画像スパム 1 通に対し，学習時に 32 ミリ秒，判定時に 150 ミリ秒のオーバヘッドがみられる．判定時は，画像情報取得と 2 度迷惑メール確率が計算されることにより，学習時に比べ，オーバヘッドが大きくなっている．この遅延は，画像スパムに対してのみであり，見逃しによるメールの移動や削除に必要な労力を考えると，許容範囲内であるといえる．

表 12　画像スパム 1 通あたりの処理時間（単位：ミリ秒）
Table 12　Processing time per image spam email (ms).

	学習時	判定時
従来方式	62	120
提案方式	94	270

6. 関 連 研 究

ベイジアンフィルタの実装方法として，POPFile[16] や Mozilla Thunderbird[17] のようにクライアント PC 上で動作するものと，bsfilter[15] や bogofilter[18]，SpamAssasin[19] のように受信サーバ上で動作するものがある．各プログラムにより，学習効果や処理時間などに違いがあるが，根本的にはすべてベイズ理論を基礎にした実装方法となっている．現在，画像スパムの画像に対して処理を行うベイズ理論に基づくプログラムは存在しない．したがって，画像スパムに対しては本文とヘッダのみで判定を行っている．

画像スパム対策としては，OCR を用いて画像からテキストを抽出するフィルタリング[20] や画像のみで分類を行う方法がある．前者は，導入のコストが高く，また画像に CAPTCHA 技術を用いた画像スパムの出現により，テキストの抽出が困難になっており，効果的なフィルタリング方法となっていない．そこで，OCR ツールを補うため，ノイズが多いテキストを発見し，それをフィルタリングに利用するアプローチが提案されている[21]．後者は，スパム画像と合法的な画像間の画像特性の違いを学習させ，その学習データを基に，新しく受信した電子メールの添付画像を分類する方法である．分類法として，SVM[22] を用いた分類[23),24] や MFoM[25] を用いた分類[26] がある．各分類法では，画像のみの情報で，スパム画像の約 80%を分類することができるが，誤検出率は最良で 5.6%であり，精度向上が課題となっている．また，テキスト処理よりも画像処理の方が時間がかかるため，分類速度が遅い．分類速度の改善法として，画像特性選択アルゴリズムが提案されている[27]．これは，分類への影響力が大きい画像特性を選択することで，利用する画像特性を減らし，画像処理を高速にする方法である．

7. ま と め

本論文では，ベイジアンフィルタにおける画像スパム対策の設計とその評価について述べた．従来のベイジアンフィルタはテキストのみに対して学習と判定を行っているため，迷惑

メールの内容を画像化する画像スパムの画像に対して対策がとれなかった．提案方式では，そこで，画像スパムの画像情報に着目し，画像情報をベイジアンフィルタに学習させ，判定に用いた対策を提案した．

GIF 画像が添付された画像スパムに関して実験を行い，従来方式に比べ，提案方式は見逃し率を下げることができることを示した．また，提案方式はテキストのみの判定結果により，画像情報を追加するかどうか決めるので，画像付きの正当な電子メールの誤検出率は，従来方式とほぼ同等である．したがって，提案方式は，従来方式以上のフィルタリング精度があるといえる．また，処理時間のオーバヘッドが小さいことも示した．

謝辞　本研究の一部は，C&C 振興財団若手研究員助成，および中島記念国際交流財団日本人若手研究者研究助成の支援を受けて行った．

参 考 文 献

1) postini: Email Monitoring + Email Filtering Blog. http://www.dicontas.co.uk/blog/quick-facts/email-spam-traffic-rockets/65/
2) フィッシング対策協議会：APWG Phishing Activity Trends Report. http://www.antiphishing.jp/report/200709-apwg-082.pdf
3) McAfee: McAfee Avert Labs Blog. http://www.avertlabs.com/research/blog/?p=170
4) Graham, P.: A Plan for Spam. http://paulgraham.com/spam.html
5) Robinson, G.: Spam Detection (2002). http://radio.weblogs.com/0101454/stories/2002/09/16/spamDetection.html
6) Robinson, G.: A Statistical Approach to the Spam Problem (2003). http://www.linuxjournal.com/article/6467
7) 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).
8) 岩永　学，田端利宏，櫻井幸一：統計フィルタリングに対する Word Salad 攻撃についての考察，コンピュータセキュリティシンポジウム 2004（CSS 2004）論文集，pp.13–17 (2004).
9) 岩永　学，田端利宏，櫻井幸一：迷惑メール内の Word Salad による統計的フィルタリングの学習データへの影響，2005 年暗号と情報セキュリティシンポジウム（SCIS 2005）予稿集，Vol.1, pp.187–192 (2005).
10) 王　戦，堀　良彰，櫻井幸一：画像スパムの外見的特徴についての考察，コンピュータセキュリティシンポジウム 2007（CSS 2007）論文集，pp.151–156 (2007).
11) CAPTCHA. http://en.wikipedia.org/wiki/CAPTCHA
12) Google. http://www.google.co.jp/
13) Wang, Z., Hori, Y. and Sakurai, K.: A Design of Image-based Spam Filtering Based On Textual and Visual Information, 情報処理学会コンピュータセキュリティ（CSEC）研究会，pp.279–284 (2008).
14) 総務省（報道資料）．http://www.soumu.go.jp/s-news/2007/070306_2.html
15) bsfilter. http://bsfilter.org/
16) POPFile. http://popfile.sourceforge.net/
17) Mozilla Thunderbird. http://www.mozilla.com/en-US/thunderbird/
18) bogofilter. http://bogofilter.sourceforge.net/
19) SpamAssasin. http://spamassassin.apache.org/
20) Barracuda. http://www.barracudanetworks.com/ns/?L=jp
21) Biggio, B., Fumera, G., Pillai, I. and Roli, F.: Image Spam Filtering by Content Obscuring Detection, 4th Conference on Email and Anti-Spam. http://www.ceas.cc/2007/papers/paper-40.pdf
22) Vapnik, V.N.: *The Nature of Statistical Learning Theory*, Springer-Verlag (1995).
23) Aradhye, H.B., Myers, G.K. and Herson, J.A.: Image Analysis for Efficient Categorization of Image-based Spam E-mail, *Proc. 8th International Conference on Document Analysis and Recognition* (*ICDAR 2005*), pp.914–918 (2005).
24) Wu, C.T., Cheng, K.T., Zhu, Q. and Wu, Y.L.: Using Visual Features For Anti-Spam Filtering, *Proc. 2005 IEEE International Conference on Image Processing* (*ICIP 2005*), pp.509–512 (2005).
25) Gao, S., Wu, W., Lee, C.H. and Chua, T.S.: A MFoM Learning Approach to Robust Multiclass Multi-Label Text Categorization, *Proc. 21st International Conference on Machine Learning* (*ICML 2004*), pp.329–336 (2004).
26) Byun, B., Lee, C.H., Webb, S. and Pu, C.: A Discriminative Classifier Learning Approach to Image Modeling and Spam Image Identification, *4th Conference on Email and Anti-Spam*. http://www.ceas.cc/2007/papers/paper-66.pdf
27) Dredze, M., Gevaryahu, R. and Bachrach, A.E.: Learning Fast Classifiers for Image Spam, *4th Conference on Email and Anti-Spam*. http://www.ceas.cc/2007/papers/paper-06.pdf

(平成 19 年 11 月 30 日受付)
(平成 20 年 6 月 3 日採録)

上村　昌裕

2007 年岡山大学工学部情報工学科卒業．現在，同大学大学院自然科学研究科博士前期課程在学中．コンピュータセキュリティに興味を持つ．

田端　利宏（正会員）

1998 年九州大学工学部情報工学科卒業．2000 年同大学大学院システム情報科学研究科修士課程修了．2002 年同大学院システム情報科学府博士後期課程修了．2001 年日本学術振興会特別研究員（DC2）．2002 年九州大学大学院システム情報科学研究院助手．2005 年岡山大学大学院自然科学研究科助教授．現在，同准教授．博士（工学）．オペレーティングシステム，コンピュータセキュリティに興味を持つ．電子情報通信学会，ACM 各会員．