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

2007－ＣＳＥＣ－38 (45)
2007／7／20

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設計と評価

上村昌裕† 田端利宏†

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，ファイルサイズ等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 Design and Evaluation of Filtering Method of Image Spam in Bayesian Filter

Masahiro Uemura† and Toshihiro Tabata†

In late years, with the spread of Internet, increase of an unwanted e-mail becomes a problem. In 2006, there is a finding that spam-mail is occupied 91% of the whole E-mail. A Bayesian filter filtering statistically with Bayes theory as an anti-unwanted e-mail measure is used widely. The filter has high precision of filtering and is able to match with the trend of an unwanted e-mail and personal preference as the characteristic. However, the image spam that does imaging of contents of an unwanted e-mail increases rapidly as the end run. Baysian filter is not able to do an appropriate learning and judgement for binary data such as images since Bayesian filter learns and judges for only text data. Therefore, in this paper, we pay attention to information, such as file size of an attached image and suggest a technique of filtering with adding these information to a corpus of existing Bayesian filter as an anti-image spam measure. In addition, we report the evaluation result.

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[1]が存在する．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，無駄な記憶領域の確保，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，これらの問題点から，電子メールの信頼性が低下している．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されている．

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とは，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させ，それらの学習データ（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[2]によると，2006年末には，画像スパムは迷惑メール全体

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

のうちの 65 %を占めている．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，本文に正当な電子メールであるかのような内容を載せ，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，ファイルサイズ等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．

本方式の利点は，画像スパムの見逃し（迷惑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，処理時間に関しては，従来のフィルタリングとほぼ変わらない時間で行える．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し，この結果により，必要であれば画像情報を加えて判定する．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，正当な電子メールであれば，ほとんどの場合，ヘッダと本文の評価で正当なメールと判断できるためである．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を迷惑メールとして誤検出（正当な電子メールを誤って迷惑メールと見なすこと）することを防止でき，判定精度を向上させることができる．

2. ベイジアンフィルタ

ベイジアンフィルタは，最初に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して，ある単語 w を含む電子メールが迷惑メールである確率 $p(w)$ を計算する．次に，この $p(w)$ を用いて判定対象の電子メール m が迷惑メールである確率 $p(m)$ を計算し，その確率が閾値を上回ったものを迷惑メールと判断する．

迷惑メール確率の計算方法として，Graham の方式[3] や Robinson の方式[4] が多く用いられている．

提案方式では，Robinson の方式を用いて判定を行った．Robinson の方式は，Graham の方式を基に提案した方式である．Robinson の方式では，単語ごとの迷惑メール確率 $f(w)$ を以下のように求める．

最初に，トークンごとの迷惑メール確率 $p(w)$ を求める．

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

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

この $p(w)$ を用いて，$f(w)$ は次のように計算される．

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

ここで，x は今まで 1 度もメールの中に出現していない単語が初めてメールに出現したときに，そのメールが迷惑メールである予測確率とし，s (strength) をその予測に与える強さとする．また，n は単語 w の出現回数とする．x と s の値は，フィルタのパフォーマンスが最適化されるように設定すべきであるが，$x = 0.5$, $s = 1$ が妥当であるとされている．

Graham の方式と比較して Robinson の方式が優れているのは，単語 w の出現回数が少ない場合をうまく扱える点である．Graham の方式では，ある単語 w がスパムメールのみに数回出現した場合，そのメールの単語の迷惑メール確率 $p(w)$ が 1 になる計算方法となっている．しかし，その程度の情報で単語 w に最大の迷惑メール確率を与えるのには問題がある．そこで，Robinson の方式では，単語 w の総出現回数が小さい場合には $p(w)$ の比重を小さくなるような計算方法をとり，まだ情報不足であるということを $f(w)$ に暗に加味することができる．それから学習が進むに従い，総出現回数 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，$n = 0$ の場合には $f(w) = x$ 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．

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

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

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

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．

これらの計算方式においては，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

るように計算される．確率が 0 に近い値は，そのトークンが，正当な電子メールに特徴的なトークンであることを表し，その電子メールは，正当な電子メールの可能性が高いことを意味する．確率が 1 に近い値は，迷惑メールに特徴的なトークンであることを表し，迷惑メールの可能性が高いことを意味する．

ベイジアンフィルタの特徴としては，判定したメールに含まれる単語を新たに学習し，出現確率のデータを更新できるという特徴がある．これにより，以後の迷惑メールの判定精度が向上していく．たとえば，迷惑メール業者が送信するメールの内容が変化するとともに，フィルタで遮断する迷惑メールの基準も変化していく．これに伴い，フィルタを使用する利用者が，受信する迷惑メールと正当な電子メールの傾向に合わせて，フィルタリング基準も変化させることができる．このため，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．

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[5]．中でも，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[2] 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．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．

テキストベースのフィルタリング，中でもベイジアンフィルタは，精度が高くて有名である．しかし，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．よって，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なければならない．

スパム送信者が送信したい内容は画像化しているので，本文が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．ヘッダは SMTP の設計上，改変を行うことができるので，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，ワードサラダのように，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く，近年増加傾向にあり，問題となっている．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 倍大きいので，

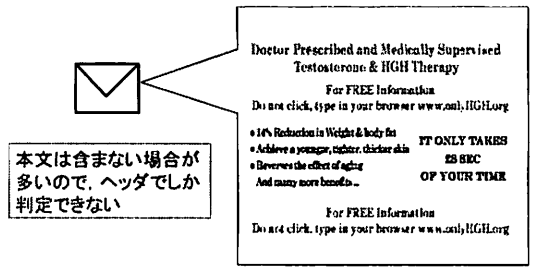

図 1 画像スパムの添付画像の例

サーバへの負担が大きいのも問題点である[2]．

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．最も古典的な手法は，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，検出をかいくぐる方法である．また，人間の目には同じように見える画像でも，機械的にはユニークなファイルを用いることで，シグネチャによる検出を困難にする方式もある．中には，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．

さらに，OCR（Optical Character Reader：光学的式文字読取装置）を用いて画像スパムをフィルタリングする技術をかいくぐるため，画像をゆがませるなどの加工を加え，人の目には読み取れる形で送信する画像スパムも登場している[2]．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．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，この結果を用いて判定させる．これにより，画像スパムのフィルタリング精度を上げることを

表 1 画像形式の内訳

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

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について検討する必要がある．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．それに比べ，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のは問題がある．

そこで，提案方式では，正当な電子メールに添付される画像のメタデータと，画像スパムの画像のメタデータの違いをコーパスに学習させ，判定を行うことにより，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を実現する．

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 によると，画像スパムは，GIF や JPEG の画像形式で画像を送信する場合が多い．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．そこで，コーパスに追加するメタデータは，ファイル名，ファイルサイズ，面積，圧縮率の 4 つとし，各情報に現れる特徴について述べる．

4.3 画像データの分析

分析に用いた画像は，表 1 で最も数が多かった 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[6] 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，各情報に表れる特徴

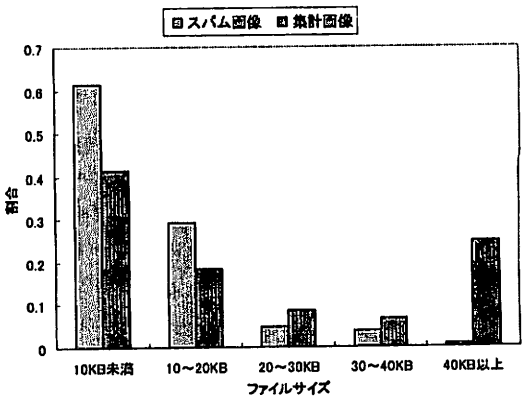

図 2 スパム画像と集計画像のファイルサイズ

について述べる．

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と考えられる．実際，同じファイル名の画像を送信する画像スパムもいくらかあった．

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．20KB 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，通常の画像よりファイルサイズが小さい場合が多い．

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．したがって，画像スパムでない画像より圧縮率が高いのではないかと考えられる．図 4 を見ると，分布にばらつきがあるが，圧縮率が 50 %以上の画像を考えると，スパム画像の方が数が多いので，集計画像よりも若干圧縮率が高いのではないかと考えられる．

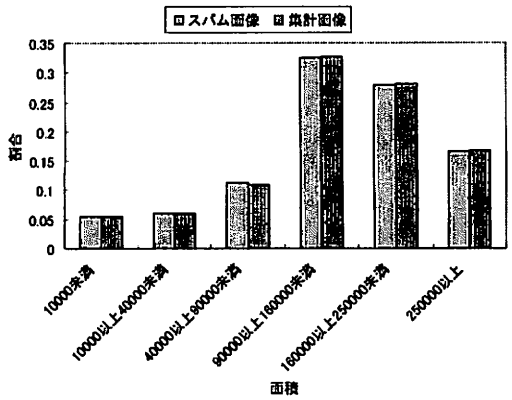

図 3　スパム画像と集計画像の面積

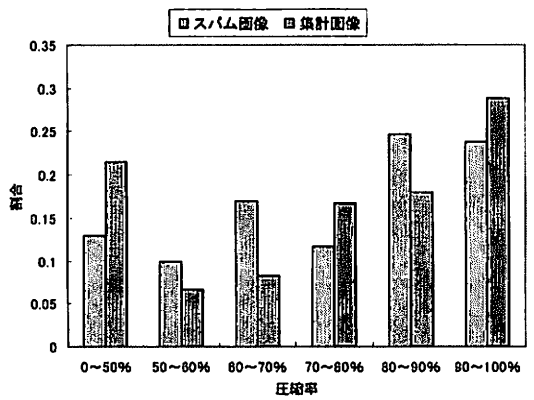

図 4　スパム画像と集計画像の圧縮率

4.4　画像データのトークン

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，コーパスに組み込む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．理由として，数値として同じ値を得る場合は少ないと考えられるので，効果的に判定に用いることができないからである．組み込む例として，ファイルサイズが 10KB～20KB の場合，“size10_20KB” といったトークンとしてコーパスに組み込む．

5. 予 備 評 価

5.1　概　　要

提案方式は，既存のベイジアンフィルタとして実装されている bsfilter[7] を基に実装した．添付画像の画像形式は，GIF と JPEG の画像形式に対応している．

表 2　実験に用いた電子メール（本文の言語は英語）

	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

本節では，画像情報組み込みによる判定を調査する実験を行う．

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 2 に示す．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かったことと，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　画像情報のコーパス組み込み方法

各画像情報は，前節での調査を基に，表 3，表 4，および表 5 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しておく．

5.4　確率計算に用いる画像情報について

今回の実験は，Robinson の方式で迷惑メール判定確率の計算を行った．Robinson の方式では，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．今回の実験では，画像情報は迷惑メールにしか現れないので，画像情報の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，いくら画像情報の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．したがって，判定に組み込む画像トークンの数を変化させて確率を計算した．

各画像情報の組み込む数を 1 個とした方式を組み込み方式 1，10 個とした方式を組み込み方式 2，30 個とした方式を組み込み方式 3，50 個とした方式を組み込み方式 4 とする．ここで，ファイル名は利用するトークン数を変化させないことにする．これは，まったく同じファイル名が現れる可能性は低く，判定時に利用するトークン数を変化させても，判定確率に与える影響が小さいと考えられるからである．

表 3　ファイルサイズのコーパス組み込み方法

ファイルサイズ	0〜10KB	10〜20KB	20〜30KB	30〜40KB	40KB 以上
トークン名	I_size10KB	I_size10_20KB	I_size20_30KB	I_size30_40KB	I_size_40KB
トークン数	98	78	12	5	8

表 4　面積のコーパス組み込み方法

面積	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 5　圧縮率のコーパス組み込み方法

圧縮率	0〜50 %	50〜60 %	60〜70 %	70〜80 %	80〜90 %	90〜100 %
トークン名	I_compress50	I_compress50_60	I_compress60_70	I_compress70_80	I_compress80_90	I_compress90_100
トークン数	27	22	18	23	50	61

表 6　判定時に利用するトークン数

	利用する各画像情報のトークン数	従来方式より増えるトークン数の合計
組み込み方式 1	1	4
組み込み方式 2	10	31
組み込み方式 3	30	91
組み込み方式 4	50	151

判定確率の計算に利用する画像情報のトークン数を表 6 に示す．たとえば組み込み方式 2 の場合，ファイルサイズのトークンを 10 個，面積のトークンを 10 個，圧縮率のトークンを 10 個，ファイル名のトークンを 1 個判定に利用するので，従来方式に比べ，判定時に利用するトークン数が 31 個増える．

5.5 実験結果

画像スパムの従来方式と提案方式の bsfilter での正当な電子メールと迷惑メールの判定確率に差はなかった．画像情報がなく，判定に用いるトークンが同じなので，当然の結果である．

判定結果を表 7 に示す．画像スパムの判定確率の各分布ごとに，判定結果として得られたメールの数と全体の割合を示した．判定確率は，0 に近いほど正当な電子メールである可能性が高く，1 に近いほど迷惑メールである可能性が高い．また，表中の “token” は，利用する各画像情報のトークン数を示す．画像スパムで判定確率が 0.5 未満のものはなかった．これは，画像スパムがヘッダと本文で，迷惑メールの特徴を持っていることを示している．しかし，ヘッダと本文だけでは，迷惑メールと断定できない画像スパムは多く存在する．

5.6 考　　察

実験結果より，組み込み方式 1 のように，判定時に用いる画像情報が少ない場合，迷惑メール判定確率の変化はあまり見られない．一方，組み込み方式 2，組み込み方式 3，組み込み方式 4 のように，判定時に用いる画像情報が多い場合，画像スパムの迷惑メール判定確率は上がっている．たとえば，閾値を 0.9 と設定した場合，従来方式の画像スパムの見逃し数は 13.5 %であるが，組み込み方式 2 では 6.5 %，組み込み方式 3 では 2.0 %，組み込み方式 4 では 0.5 %と大幅に見逃しが減っていることがわかる．

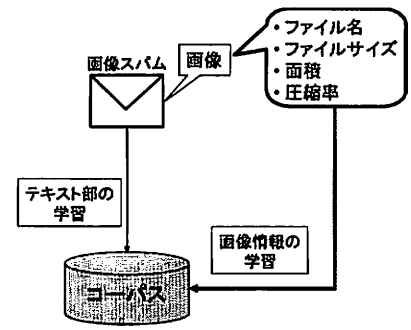

図 5　学習時の流れ

見逃し数は減少するが，このままでは，正当な電子メールに画像が添付されていたときに与える影響も大きくなることが考えられる．

しかし，画像付きの正当な電子メールの場合，ヘッダや本文に正当な電子メールであることを示す十分な情報が含まれていることが期待できる．このため，ヘッダと本文で正当な電子メールと判断できるメールの場合，画像スパムのコーパスを加えて判定する必要はない．

6. 実装と評価

6.1 実装方式

6.1.1 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，ファイル名等の画像情報をコーパスに学習させるようにする．

表 7　画像トークン数追加による画像スパムの判定結果

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	6	4	0	25	148
	(8.5 %)	(3.0 %)	(2.0 %)	(0.0 %)	(12.5 %)	(74.0 %)
組み込み方式 1	15	2	5	5	22	151
(token=1)	(7.5 %)	(1.0 %)	(2.5 %)	(2.5 %)	(11.0 %)	(75.5 %)
組み込み方式 2	10	1	1	1	25	162
(token=10)	(5.0 %)	(0.5 %)	(0.5 %)	(0.5 %)	(12.5 %)	(81.0 %)
組み込み方式 3	3	0	3	1	7	186
(token=30)	(1.5 %)	(0.0 %)	(1.5 %)	(0.5 %)	(3.5 %)	(93.0 %)
組み込み方式 4	0	1	0	0	8	191
(token=50)	(0.0 %)	(0.5 %)	(0.0 %)	(0.0 %)	(4.0 %)	(95.5 %)

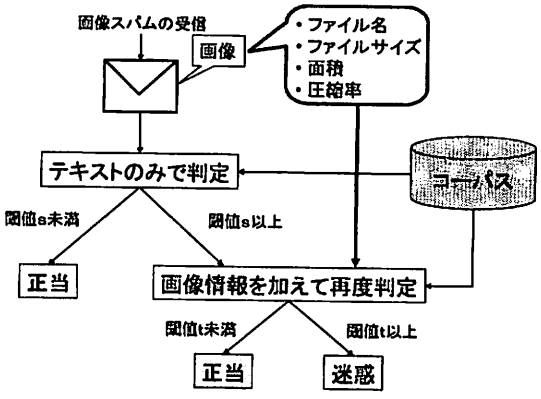

図 6　判定時の流れ

6.1.2 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 点の傾向がみられることが多い．

- 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．
- 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．

これらの傾向を基に，画像が添付されたメールについて，以下の手順でフィルタリングを行う方式を提案し，実装する．

(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 以上の場合，画像情報も加えて判定する提案方式で迷惑メール確率を計算する．

(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．

判定時の流れを図 6 に示す．閾値 s は 0.4 と設定し，画像が添付されたメールの迷惑メール確率が 0.4 以上の場合，確率が再計算される．また，テキストのみでの判定確率が 0.9 以上の場合，そのメールは画像の有無に関係なくほぼスパムであると考えられるので，再計算は行わなかった．本実験では，判定に用いる画像トークン数は，テキストのみで判定した時の全トークン数の割り合いで変化させ，実験を行った．例えば，テキストのみで判定した時の全トークン数が 100 で，判定に用いる画像トークン数を 10 %とした時，提案方式の判定時に用いられる全トークン数は 110（ファイル名:1，その他:各 3）である．

6.2 評　　価

学習と判定に用いたメールは，表 2 と同じメールセットとした．判定結果を表 8 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 %，9.5 %，3.5 %，0.5 %である．画像トークン数の 10 %追加では，従来方式と比べ，4 %の改善が見られたが，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が，画像情報の影響をうけやすく，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．

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，適度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

表 8 提案方式の画像スパムの判定結果

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 %)

表 9 画像スパム 1 通あたりの処理時間（単位：ミリ秒）

	判定時	学習時
従来方式	62	120
判定方式	94	270

うので，従来方式と誤検出率はほぼ変わらないといえる．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．結果より，閾値を 0.8 とすると，見逃し率が低くなることがわかる．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．

処理時間に関しては，Pentium III(1.26GHz) 搭載の計算機を用いて測定を行った．結果を表 9 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 ミリ秒遅くなり，判定時に 150 ミリ秒遅くなった．この程度の遅延は，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．

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，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たので，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．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，判定に用いることによる対策を提案した．次に，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が下がることを示した．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，画像スパムに対しての精度は，従来方式以上であるといえる．

今後の課題として，他の画像形式（JPEG，PNG 等）への拡張，画像情報のより効果的なコーパス組み込み方法がある．

謝辞 本研究の一部は，C & C 振興財団 若手研究員助成，及び科学技術振興機構 戦略的国際科学技術協力推進事業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) McAfee : McAfee Avert Labs Blog. http://www.avertlabs.com/research/blog/?p=170
3) Graham, P.: A Plan for Spam. http://paulgraham.com/spam.html
4) Robinson, G.: Spam Detection. http://radio.weblogs.com/0101454/stories/2002/09/16/spamDetection.html
5) 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.
6) Google : http://www.google.co.jp/
7) bsfilter : http://bsfilter.org/