機械学習によるメールヘッダ情報に基づく迷惑メールフィルタリング

杉井 学[†1] 角 朝香[*1,†2] 松野浩嗣[†1,†2]

迷惑メールのフィルタリングには，さまざまな手法とアプリケーションソフトが開発されている．我々は機械学習システムを用いてメールヘッダ情報から特徴を抽出してフィルタリングに利用する新たな手法を開発した．これまでに，メールに含まれる単語の出現頻度とその語順を組み合わせた情報を，機械学習システムで解析し，メールの特徴を決定木で表現して利用する手法を確立した．同様の手法を用いて，メールヘッダ情報からフィルタリングに必要な特徴を抽出して評価したところ，正規メールのヘッダに着目した特徴解析によって高い精度で迷惑メールの分類を行えることが明らかになった．

A Novel Email Filtering Method Based on the Information in Email Header with the Use of a Machine Learning Technique

Manabu Sugii,[†1] Sayaka Kado[*1,†2] and Hiroshi Matsuno[†1,†2]

A lots of methods and application software had been developed for spam mail filtering. We have developed a new filtering system to classify spam and non-spam mails by the use of a rule extracted from email headers by a machine learning system. Previously, we had established a classification method based on the appearance order and the frequency of a word. This rule was represented by a decision tree producted from the machine learning system. In this study, we evaluated the rule extracted from email headers by the same classification method, with which could classify emails with a high accuracy rate by focusing on non-spam mail headers.

1. はじめに

インターネット利用者の増加や携帯電話の普及に伴い，メールは仕事でも日常生活でも必要不可欠なものとなった．そのため，流通するメールの９割近くが迷惑メールになっている現状は，さまざまところで不具合を生じ社会的問題に発展している[1)]．もはや，日々送りつけられる迷惑メールをユーザが手作業で取り除くことは，現実的ではない状態に陥っている．

このため，様々な迷惑メール対策手法が実現化されており[2)3)]，迷惑メールの特長の変化によって対策手法も移り変わってきた[4)]．近年ではベイズ理論を応用した方法が採用され，迷惑メールの高い分類精度を実現している．しかし迷惑メール送信者は，これらの手法をかいくぐる新たな方法を次々に開発し[1)]，さらに大量の迷惑メールを送信してくることから，迷惑メール送信者とメールサーバ管理者の攻防は「イタチごっこ」の様相を呈している．

そこで，我々の提案するメールフィルタでは，個々のメール利用者が受信する正規メールの特徴抽出に目を向けて分類する手法を採用してきた[5)]．つまりこの手法は，迷惑メールの特徴を利用しつつも，正規メールの特徴を抽出して活用することで，どんなに迷惑メールが多様化し複雑化したとしても，その影響を受けることのない分類を可能にする手法である．また，機械学習システムを用いてメールヘッダ情報から特徴を抽出して分類に用いるだけで，十分実用可能な分類精度を得られることが明らかとなった．

2. 機械学習システムによるメール特徴の抽出と学習領域の検討

我々はこれまで，メールを構成する単語の出現頻度とその語順を学習要素にして，機械学習システムで作成した決定木をメールの分類に利用してきた．これは，文章の内容を表す特徴は，文章中に存在するいくつかの重要単語で表現され，これら重要単語は同様の内容を示す文章内には高い頻度で出現することを利用した方法である．また我々は，従来のフィルタリング手法に用いられる単語の出現頻度情報だけではなく，それらの出現順も学習要素に加えている．

機械学習システムには，BONSAI を用いた．BONSAI は，概念学習の一つである確率的

†1 山口大学大学情報機構メディア基盤センター
Media and Information Technology Center, Yamaguchi University
†2 山口大学大学院理工学研究科
Graduate School of Science and Engineering, Yamaguchi University
∗1 現在，日本ラッド株式会社
Presently with Nippon RAD Inc.

近似学習（Probably Approximately Correct Learning: PAC 学習）と呼ばれる学習パラダイムに基づき開発されたシステムである[6]．このシステムは，一次元の記号列データを対象にし，二つの学習例から，それらを区別する仮説を決定木として提示する機能を持つ．さらに，決定木を構成する各ノードは，indexing と呼ばれる要素のグルーピングによって，学習例として与えられた一次元の記号列データの要素数を減らして最適化した別の記号列で表現される．

これまでに，これらの手法を用いて高い精度でのメールフィルタを実現しているが[5)7)]，得られたメール分類特徴を表す決定木のパターンから，正規メールのヘッダ情報に，より強い特徴が含まれていることが示唆された．そこで，BONSAI に入力する学習例の学習領域をメールヘッダおよびメール本文に分割し，分類精度を比較することで，最も分類効率の高い学習領域（強いメール分類特徴を持つ領域）の検討を行った．

3. BONSAI を利用したメールフィルタリング手順

図 1a は，BONSAI を用いた解析手法の中核をなす文字列変換プロセスを説明するための一例である．メールに含まれる文字列は，文節または単語に分割され，それらの出現頻度にしたがったアルファベットに変換される．その後，機械学習システムが行う indexing と呼ばれる要素のグルーピングによって，文字列の出現頻度と出現順の特徴を維持したまま，機械学習システムで用いる新たな記号列に変換される．

また，図 1b は，メールの文章特徴を BONSAI の出力する決定木で表現して分類に利用するための一連のプロセスを模式的に表したものである．学習プロセスでは，メールに含まれる文字列を文節または単語に分解する処理，それぞれの文節または単語の出現頻度を算出する処理，出現頻度に従ったアルファベットへの変換処理，BONSAI による indexing と決定木作成処理などで構成される．分類対象となるメールの分類は，学習例の場合と同様に，文節または単語に分解する処理，それぞれの文節または単語の出現頻度を算出する処理，出現頻度に従ったアルファベットへの変換処理などで構成されるが，学習段階で BONSAI が作成した indexing 結果を用いた記号列への変換処理や BONSAI が作成した決定木を使い，決定木を構成するノードパターンが，完全一致の状態で記号列変換した分類対象メールの中に含まれるか含まれないかを判断する処理などが含まれる．

3.1 “分かち書き” 処理（文節または単語への分解）

学習例に用いるメールは，メールヘッダを含むすべての情報を，Extended UNIX Code（EUC-JP）に統一した．その後，メールヘッダまたはメール本文に領域を分割し，それら

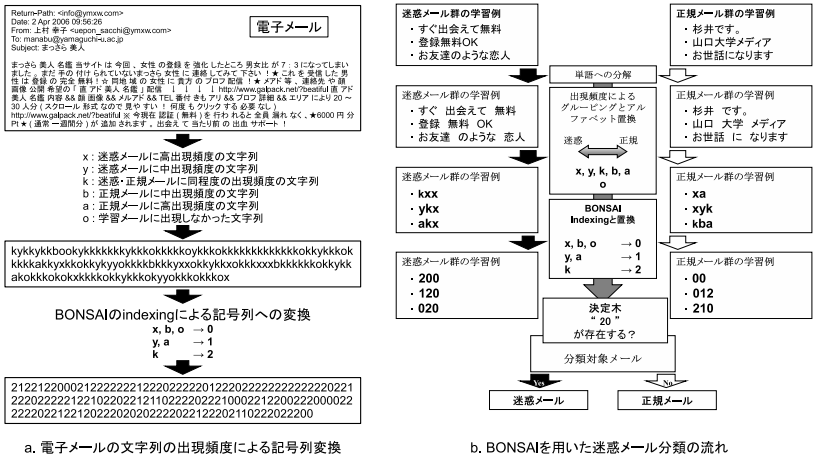

図 1 BONSAI を用いたスパムメール分類の流れ
Fig. 1 Process of spam mail filtering system with BONSAI.

に含まれる文節または単語単位の文字列の出現頻度を算出するために，KAKASI（2.3.4）[8)] を用いて半角スペースをデリミタとする分かち書きされた文章へと変換した．文節または単語への分割が言語として正しく行われない場合があるが，我々の採用した手法では言語の構造や意味を理解した処理ではなく，特定の文字列の出現頻度と出現順のみを利用することから，言語的な処理のエラーは考慮しなかった．この他，KAKASI による分かち書きと分割された文字列の出現頻度の算出をより正確にするために，前処理として漢字コードの変換直後に，下記の処理を行った．

- 特定の記号の除去（表 1）．
- アルファベット大文字の小文字への変換
- 英単語の活用形を現在形に変換

3.2 文字列の出現頻度の算出

前節の “分かち書き” 処理で分割したすべての文字列について，学習例全体での出現総数と，迷惑，正規それぞれのメール群での出現総数を算出した．迷惑，正規それぞれのメール群での出現総数の違いを補正するため，式 (1) の r を用いて式 (2), (3) のように抽出文字列数の補正を行った．式 (4) は，学習例全体での文字列数を表している．

表 1 取り除く記号の一覧
Table 1 The list of removed characters.

半角記号	() [] {} | ; : ^ < > , . ? ” ’
全角記号	． ． ， ， （） 「」 ” ’

$$r = \frac{\text{迷惑メール群での抽出文字列総数}}{\text{正規メール群での抽出文字列総数}} \tag{1}$$

$$N_s(word) = \text{文字列 } word \text{ の迷惑メール群における出現数} \tag{2}$$

$$N_h(word) = \text{文字列 } word \text{ の正規メール群における出現数} \times \text{r} \tag{3}$$

$$N_{all}(word) = N_s(word) + N_h(word) \tag{4}$$

次に，ある文字列の学習例全体での出現率，迷惑メール群での出現率，正規メール群での出現率をそれぞれ求めた．式 (5), (6), (7) は各々，文字列 $word$ に対する全体の出現率 $rate(word)$，迷惑メール群での文字列 $word$ の出現頻度 $freq_s(word)$，正規メール群での文字列 $word$ の出現頻度 $freq_h(word)$ を表している．

$$rate(word) = \log(N_{all}(word)) \tag{5}$$

$$freq_s(word) = \frac{N_s(word)}{N_{all}(word)} \tag{6}$$

$$freq_h(word) = \frac{N_h(word)}{N_{all}(word)} \tag{7}$$

3.3 出現頻度による記号列変換

学習例として用いたメール群において，迷惑メール群と正規メール群各々について特徴的に高い出現頻度を示す文字列を定義した．つまり，式 (5) で算出した出現率 $rate(word)$ が，式 (8) で表す基準値 $rate\_base$ 以上の値を示す文字列 $word$ について，表 2 に示すように，その出現率 $freq_s(word)$ または $freq_h(word)$ の値に従って，5 つの記号からなるアルファベット {x, y, a, b, k} に置換した．なお，学習例のメール群に含まれるすべての文字列の中での出現率の最大値を $rate\_max$ とした．

$$rate\_base = \frac{rate\_max}{2} - 1 \tag{8}$$

図 2 は，縦軸に式 (6) で算出した迷惑メール群での文字列 $word$ の出現頻度 $freq_s(word)$ を，横軸に式 (5) で算出した文字列 $word$ に対する全体の出現率 $rate(word)$ を配置して，メール文章中に含まれる文字列を一つの点としてプロットした場合の予想散点分布図である．また，表 2 で定義した出現頻度を表すアルファベットに変換される文字列の分布域を模式的に表したものである．x，y に置き換えられた文字列は，迷惑メール群に特徴的に現れ，a，b に置き換えれた文字列は，正規メール群に特徴的に現れる．k に置き換えられた文字列は迷惑メール群，正規メール群の両メール群に一定の頻度で現れる文字列である．また，x，y，a，b，k のいずれにも置き換えられなかった文字列，つまり出現頻度が式 (8) で表す基準値 $rate\_base$ 以下の値を示す文字列や学習例中に存在しなかった文字列については，学習メール群において強い特徴を持つ文字列ではないとみなし，o で置き換えた．この変換により，学習例に含まれるメールは，6 つの記号からなるアルファベット {x，y，a，b，k，

表 2 出現頻度による記号列変換表
Table 2 Conversion of a word based on an appearance frequency of the word.

変換記号	$freq_s(word)$	$freq_h(word)$
x	0.8 以上	-
y	0.8 未満，0.6 以上	-
k	0.6 未満，0.4 以上	0.6 未満，0.4 以上
b	-	0.8 未満，0.6 以上
a	-	0.8 以上

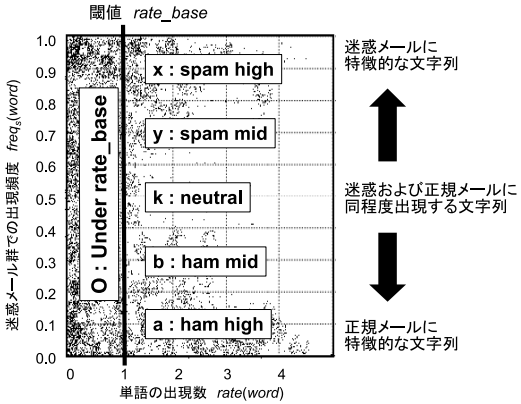

図 2 文字列の出現頻度と変換記号の関係
Fig. 2 The relationship between appearance frequencies and converted symbols of words.

o} で表現される．

3.4 BONSAI によるメール特徴の学習と分類

出現頻度による記号列変換によって 6 つの記号からなるアルファベットに変換されたメール情報を，正規メール群と，迷惑メール群の二つの学習例として BONSAI に入力した．BONSAI は，6 つの記号からなるアルファベットを indexing し，二つの学習例を最も効率よく分類できる規則を決定木として出力する．決定木は，出現頻度を表すアルファベット {x，y，a，b，k，o} を indexing して得られた indexing 記号の配列をノードに持ち（図 3），各ノードのパターンは学習例と同様に記号列変換された分類対象メール中に，完全一致の状態で存在するか存在しないかの指標に用られる．

すなわち図 1b で示す一連のプロセスを模式的に表した例の場合，迷惑メールに含まれる “すぐ出会えて無料” という文字列は，“kxx” に変換される．その後，BONSAI による indexing によって，アルファベットの 6 つの記号は，“x, b, o” が 0 に，“y, a” が 1 に，“k” が 2 のように indexing 記号に置き換えられ，先の “kxx” は “200” に変換される．最後に，indexing 記号に変換された記号列中から，BONSAI は一定の法則を見つけ出し，決定木として出力する．この例の場合，決定木を構成するノード “20” が存在すれば迷惑メール，存在しなければ正規メールに分類されることになる．

4. メール特徴の決定木表現

BONSAI が出力する決定木は，indexing 記号列で表される複数のノードで構成され，そのノードパターンが迷惑メール群に特徴的に出現頻度の高い文字列で構成されるのか，正規メール群に特徴的に出現頻度の高い文字列で構成されるのかを確認できる．以下に，BONSAI の提示した indexing および決定木の表す特徴を示す．

4.1 メールサンプルデータ

実験には，2007 TREC Public Spam Corpus（TREC07）[9] の partial データセットを用いた．TREC07 の partial データセットは 2007 年 4 月 8 日から 2007 年 7 月 6 日のメール 30,338 通（迷惑メール 6,280 通，正規メール 24,058 通）から構成される．メールは時系列順に収められており，本解析を行う際も時系列順に取り扱った．Partial データセットの時系列の先頭から迷惑メール，正規メールそれぞれ 1,000 通ずつ選出し，学習用メール群とした．また，残りの 28,338 通（迷惑メール 5,280 通，正規メール 23,058 通）を分類対象メール群とし，得られた決定木の精度評価に用いた．

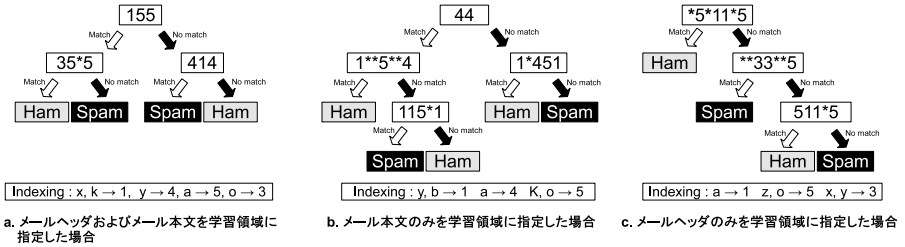

図 3 BONSAI が出力したメール判定の決定木
Fig. 3 Decision trees obtaind from BONSAI to classify emails.

4.2 BONSAI による indexing および決定木の特徴

図 3 は，学習例を構成するアルファベット {x，y，a，b，k，o} の 6 つの記号に対して，BONSAI が indexing を行い，学習例である迷惑メール群と正規メール群を最も効率よく分類できる規則を決定木として出力した結果である．図 3a は，学習例の学習領域としてメールヘッダおよびメール本文両方を指定，図 3b は，学習領域としてメール本文のみを指定，図 3c は，学習領域としてメールヘッダのみを指定したときに出力された決定木を示している．BONSAI の学習パラメータとして，Indexing Size を 6，Max Pattern Length を 7 に設定した．Indexing Size は，学習例に含まれる記号に対して，indexing を行う際の要素の最大分類数，Max Pattern Length は，決定木を構成しているノードのパターンの最大長を定義している．これらのパラメータの調整により，より精度を高めることができる可能性はあるが，BONSAI による indexing と決定木作成の処理時間を考慮し，最適化は今後の課題とした．

図 3a で BONSAI は，学習例を構成するアルファベット {x，y，a，b，k，o} を indexing によって x と k を 1，y を 4，a を 5，o を 3 と分類した．残りの b は*で表現された部分のみで許容される．*は Insignificant indexing symbol であり[10]，学習例を構成するアルファベット {x，y，a，b，k，o} すべての記号がノードパターン中の*部分で許容されることを意味する．つまり，Insignificant indexing symbol で表現された部分のみで許容される記号は，BONSAI が indexing 処理を行う際に，ノードパターンを構成する重要な位置には関与しない，あるいは特徴配列には関与しない記号であることを意味している．図 3a の決定木のノードパターン 35*5 を例にとると，o に始まり，続いて a，次にすべての記号 {x, y, a, b, k, o} のうちの一つ，最後に a で終わる記号列を意味しており，正規メール群に特徴的に

表 3 学習領域の違いによるメール分類精度の比較
Table 3 Comparison of an accuracy rate of the classification with the use of a different mail field.

学習数	1000			100			20		
学習領域	全文	本文	ヘッダ	全文	本文	ヘッダ	全文	本文	ヘッダ
分類対象数	28338								
Spam 数	5280								
Ham 数	23058								
正 Spam	5076	3892	5140	5009	3733	4907	4525	3012	4604
誤 Spam	204	1388	140	271	1547	373	755	2268	676
正 Ham	22305	20034	22506	22199	19827	22549	22459	19763	22168
誤 Ham	753	3024	552	859	3231	509	599	3295	890
FP	0.0326	0.1311	0.0239	0.0372	0.1401	0.0220	0.0259	0.1429	0.0385
FN	0.0386	0.2628	0.0265	0.0513	0.2929	0.0706	0.1429	0.4295	0.1280
TE	0.0337	0.1556	0.0244	0.0398	0.1686	0.0311	0.0477	0.1963	0.0552

出現する文字列で構成される特徴パターンを表している．

図 3b および c も同様に，いずれも正規メール群に特徴的に出現する文字列と迷惑メールに特徴的に出現する文字列両方をノードに配置した決定木となっている．しかし，注目すべきはいずれの決定木も根ノードのパターンは正規メールに特徴的にみられる文字列で構成されており，決定木作成の際に正規メール群に強い特徴を見出していると言える．

ノードの数とその構成を比較すると，メール本文のみを学習領域とした場合に，若干ノード数およびその構成が複雑化した．受信日時や受信者の異なるメールを学習例として同一条件で繰り返し決定木を作成しても，同様の傾向が見られた．これは，学習領域の文字列構成が複雑であるために，決定木を構成するノードパターンの抽出が困難になり，その結果ノード構成を複雑化して分類精度を確保したと考えられる．これまでの報告[7] では，学習領域にメールヘッダおよびメール本文両方の領域を用いた場合，決定木を構成するノードパターンはメールヘッダ領域から多く抽出されることがわかっている．

5. 学習領域および学習例数の変化と分類精度の比較

4.1 で示したメールサンプルデータを用い，学習例数と学習領域を変化させて出力される決定木の分類精度を比較した．分類結果を表 3 に示す．

学習領域は BONSAI に投入するメールの学習領域を表し，全文はメールヘッダおよびメール本文両方の文字列，本文はメール本文の文字列のみ，ヘッダはメールヘッダの文字列のみを解析に用いたことを表している．表 3 中の Spam は迷惑メール，Ham は正規メールを示す．正 Spam（Ham）は迷惑メールまたは正規メールを正しく分類したメール数である．誤 Spam（Ham）は，迷惑メールを正規メールに，または正規メールを迷惑メールに誤分類したメール数である．FP（*False Positive*）は，正規メールを迷惑メールに誤分類した割合を示し，FN（*False Negative*）は，迷惑メールを正規メールに誤分類した割合を示し，式 (9)，(10) から算出した．また，TE（*Total Error*）は，分類対象メール全体で誤分類した割合であり，式 (11) から算出した．

$$FalsePositive = \frac{\text{正規メール誤分類数}}{\text{分類対象の正規メール数}} \tag{9}$$

$$FalseNegative = \frac{\text{迷惑メール誤分類数}}{\text{分類対象の迷惑メール数}} \tag{10}$$

$$TotalError = \frac{\text{分類対象メール全体での誤分類数}}{\text{分類対象メール数}} \tag{11}$$

False Positive，*False Negative*，*Total Error* それぞれの値が低いほどフィルタリングシステムの分類精度は高くなる．*False Positive* と *False Negative* はトレード・オフの関係にあるが，正規メールの誤分類はユーザに多大な損害を与えるため，一般に *False Positive* の値の低さが *False Negative* の値の低さより重要視される．

図 4 に表 3 を元にした学習用メール数（20, 100, 1,000）に対する *False Positive*，*False Negative*，*Total Error* の変化を表す．学習領域の違いにかかわらず，学習例数が増加すれば，*False Positive*，*False Negative*，*Total Error* それぞれの値が概ね低下しており，フィルタリング性能が向上している．ただし，100 以上の学習例数でのそれぞれの値には大きな変化はなく，正規および迷惑メールそれぞれ 100 通程度のメールを学習に用いれば，十分な学習例が確保できていることが分かった（図 4）．

学習領域の違いによる，フィルタリング性能の違いは明確で，メールヘッダのみを学習領域とした場合とメールヘッダおよびメール本文両方を学習領域とした場合の性能が高いことが明らかになった．メールヘッダのみを学習領域とした場合は，メール本文のみの場合に比べて正規メールの回収率にして 10%以上性能が向上している．言い換えれば，今回の条件では，BONSAI が抽出することができるメール本文領域に共通する特徴が存在しなかったために，メールヘッダとメール本文両方を学習領域として情報量を増やしても，分類精度の向上にはつながらず，むしろノードパターンと偶然一致する文字列を含むノイズ領域となり，若干の精度の低下につながったと推測している．

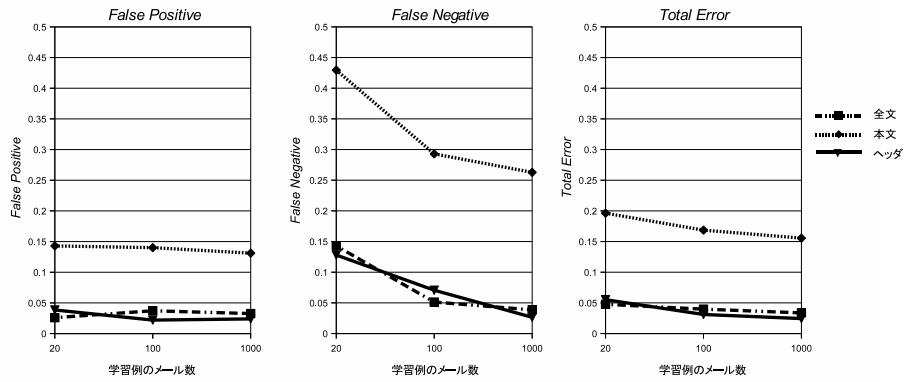

図 4 学習領域の違いによるメール分類精度の変化
Fig. 4 Performance shift of the classification with the use of a different mail field.

6. 考　　察

我々の提案する手法では，正規メールの特徴に注目し，またメールヘッダ情報を利用することで，正規メールの特徴に重点を置く高い精度のフィルタリングができることを明らかにした．このメカニズムは，BONSAI の作成する決定木の根ノードのほとんどが正規メールから抽出される特徴で構成されていることや，図 4 で示す *False Positive* つまり迷惑メールを正規メールに間違える確率が低いことからも推測できる．

また，メール本文領域よりむしろメールヘッダ領域に共通する特徴を抽出して分類に利用していることが明らかとなり，メールアドレスや中継サーバなどの情報を利用していると考えられる．これまでの報告[7] でも，決定木を構成するノードパターンで多く抽出される文字列の中に，メールアドレスがある．学習領域としてメールヘッダを与えられた BONSAI は，出現頻度の高いメールアドレスや中継サーバなどのリストに近い規則を作っている可能性が高い．また，メールヘッダ領域には書式が定められていることを考慮すれば，決められた書式に当てはまらない迷惑メールが分別されている可能性も高い．つまり，メールヘッダに含まれる文字列の出現頻度を利用することで，メール中継サーバデータベースを参照することと同様の効果をもたらし，文字列の出現順を利用することで，メールヘッダ情報の整合性を検証する方法と類似の効果をもたらしていると考えている．

7. おわりに

本稿では，どんなに迷惑メールが多様化してもその影響を受けることのない手法を追求した．それは，迷惑メールの特徴だけでなく，むしろ正規メールの特徴を用いて分類することで実現できると考えている．この方法は，受信する正規メールの特徴を機械学習を使ってユーザごとに抽出する必要があるものの，既存のフィルタリング手法をかいくぐるために日々変化していく迷惑メールの影響を受けることなく，高い精度でメールフィルタリングを実現できる．BONSAI が，正規メールと迷惑メールを分類する規則として，メールヘッダからどのように何を抽出しているかは，今後の調査課題として残っているが，複雑な言語情報から特徴を抽出する場合よりも，一定の規則に従って記述されたメールヘッダ情報から特徴を抽出するほうが，より小さく些細な変化をも捕えることができるのではないかと考えている．また実用化段階においては，分類精度を維持するために，誤分類されたメールを，その都度新たな学習例として再学習させる仕組みが必要である．

謝辞　本研究の一部は，科学研究費補助金若手研究（B）(21700078）の援助を受けている．

参　考　文　献

1) 景山 忠史, “spam メールの現状”，情報処理, Vol.46, No.7, pp.747-751, 2005.
2) Graham, P., “A Plan For Spam”: http://www.paulgraham.com/spam.html
3) Graham, P., “Better Bayesian Filtering”: http://www.paulgraham.com/better.html
4) 安東 孝二, “世界の電子メールを spam 制御へ”，情報処理, Vol.46, No.7, pp.741-746, 2005.
5) 杉井 学，松野 浩嗣，“機械学習によるスパムメールの特徴の決定木表現”，情報処理学会研究報告 2007-DPS-130(16):pp.183-188，2007.
6) Shimozono, S., Shinohara, A., Shinohara, T., Miyano, S., Kuhara, S., and Arikawa, S., “Knowledge Acquisition from Amino Acid Sequences by Machine Learning System BONSAI”, *Trans. Inform. Process. Soc. Japan*, Vol.35, pp.2009-2018, 1994.
7) 角　朝香，杉井　学，松野　浩嗣，“機械学習を応用したスパムメールフィルタリング手法の検討と評価”，マルチメディア通信と分散処理ワークショップ論文集，pp.201-204, 2009.
8) KAKASI-漢字→かな (ローマ字変換) プログラム: http://kakasi.namazu.org/
9) TREC 2007 Public Corpus: http://plg.uwaterloo.ca/g̃vcormac/treccorpus07/
10) Sugii，M., Okada, R., Matsuno, H., Miyano, S., “Performance Improvement in Protein *N*-Myristoyl Classification by BONSAI with Insignificant Indexing Symbol”, *Genome Informatics*, Vol.32, pp.277-286, 2007.