# スパムフィルタリングのための構造解析手法の提案

小林　孝史[*1]　辻野　文音[*1]　佐々木　梨絵[*2]
前田　利之[*3]　荒井　良徳[*4]　井上　敦[*5]　林　勲[*1]

## A proposal of structure analysis method for spam filtering

Takashi Kobayashi[*1], Ayane Tsujino[*1], Rie Sasaki[*2],
Toshiyuki Maeda[*3], Yoshinori Arai[*4], Atsushi Inoue[*5], Isao Hayashi[*1]

**Abstract** – **The development of the Internet has become a widespread social problem of spam. As a result, many mail user agent software are developed that have the function of spam filtering. However, in these filters, they only have the probabilistic selection methods, and the distinction methods are not based on the kind and structure of spam mail.**
**In this paper, we propose an analysis model that estimate the distinction class of spam mail by various input properties. These properties are picked out by extracting the header fields and body text of spam mails. Then, we classify time-seriesed mail dataset by using data mining methods, and verify the classification by n-fold cross validation method. Finally, we make a study of its usefulness by actually received mail.**

**Keywords** : **スパムメールの判別, 構造同定, データマイニング, クロスバリデーション, 時期的変化**

## 1.　はじめに

スパムメールとは，一般に広告・宣伝・勧誘・誘導・詐欺を目的とする電子メールを使ったメッセージで，不特定多数の相手に自動的に送り付けられるものである．また，受信者の意向を無視して送られてくるメッセージや，コンピュータウイルスやワームの動作によって無差別に発信されるメッセージも含まれる [1]．スパムメールは送信される全メールの 97%にも上っており [2]，インターネットの回線やメールサーバの資源の浪費に繋がっている．スパムメールが大量に送信されていたとしても，望まないメールを受信者は受け取る必要はない．

従来のスパム判別の方法等では，受信したメールをスパムメールや非スパムメールに確率的に分類する判断基準に頼っており，スパムメールや非スパムメールが同じクラスに属するという矛盾を内包している可能性もあり，それが識別率を低下させている原因にもなっている．また，フィルタリングのためのクラス判別であったり，さまざまな属性の組み合わせを最適化することにより識別率を向上させることに留まっており，スパムメールの構造的な把握には至っていない．

本論文では，メールヘッダの各要素からスパムメールの特徴を抽出し，それを利用した判別方法を検討し，実際に受信したメールを用いてその有効性を検証する．判別モデルの検証方法として 3-fold cross validation を用い，それぞれのデータセットから得られる判別モデルの安定性を確認する．これらを通じてスパムメールの構造を同定し，よりよいスパムフィルタリングへの応用を検討する．

## 2.　関連研究と本研究の有用性

スパムメールの識別に関する研究は，代表的なものとして 3 つのタイプがある．識別率の向上および大規模化する研究，クラス判別のための属性の抽出に関する研究，そして，スパムメールを判別するための特徴抽出とその最適化である．

Yong Hu らの研究 [5] では，ヘッダーのみを解析することによってスパムメールを検出するフレームワークを提案している．Lee ら [3] は，抽出された多数の特徴の組み合わせを最適化することにより識別率の向上を図っている．Almeida らの研究 [4] では，スパムメールを分類するための手法を改良し，より良いとされるスパムメールのフィルタリングの評価指標と共に提案している．Soranamageswari らの研究 [6] では，ニューラルネットワークを用いた画像スパムメールの分類のための特徴抽出手法について提案している．これらの

*1: 関西大学　総合情報学部
*2: エクストランス株式会社
*3: 阪南大学　経営情報学部
*4: 東京工芸大学　工学部
*5: Department of Computer Science, Eastern Washington University
*1: Faculty of Informatics, Kansai University
*2: X-Trans Inc.
*3: Faculty of Management Information, Hannan University
*4: Faculty of Engineering, Tokyo Polytechnic University
*5: Department of Computer Science, Eastern Washington University

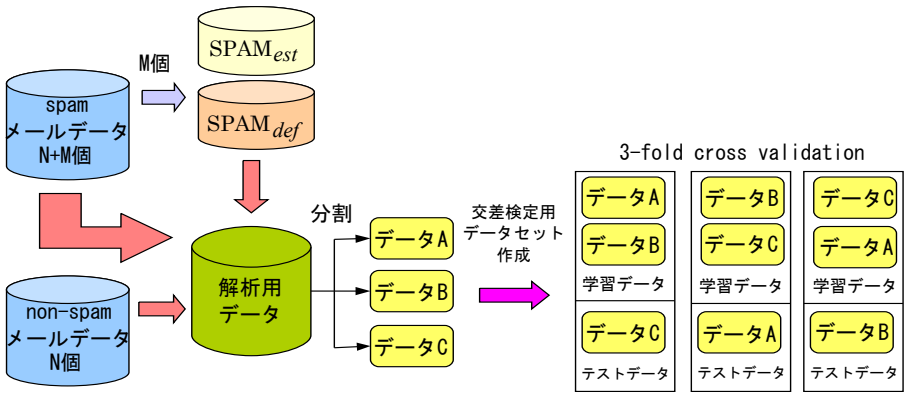


図 1　解析データの作成手順
Fig. 1　composition of analysis datasets

既存研究においては，従来の手法の改良や新しい特徴の抽出方法についての研究が主なものである．

一方，これまでに，我々は無線 LAN のアクセスポイントの最適配置の問題に Fuzzy ID3 を判別ルールの獲得に適用しており [7]，それ以外のスパムメール等の識別への応用を検討してきた．本研究での手法では，スパムメールを解析することによって後述するような属性を抽出するだけでなく，スパムメールと非スパムメールを判別するための属性がどのように利用されているか，それらの属性が判別にどの程度重要なものになっているか，ということを明らかにすることができる．また，継続的に分析することにより，1 年程度の期間の中での構造の時間的な変化を捉えることができる．既存研究のようなデータマイニングにはない分析手法と過去の時間的な傾向分析を加えたスパムメールの識別，フィルタリング手法として有益なものであると考えている．これらは過去に受信したメールを分析した結果として捉えることができる特徴であるが，この手法を応用することにより，未来のスパムメールについての傾向分析や，到着時間などの予測に役立つと考えられる．

## 3.　解析データの作成手順と解析手順

解析データの作成手順を図 1 に示す．過去 15ヶ月分 5 セットのメールデータとして，N+M 個のスパムメールデータと N 個の非スパムメールデータを用いる．まず，スパムメールデータと非スパムデータの差分である M 個をスパムメールデータからランダムに選出する．選出されたスパムメールの本文を単語に分割して，各々の単語の出現回数をカウントし，降順に並べる．そして，出現回数の多い単語を集めた $SPAM_{est}$ と出現回数が 1 度の単語を集めた $SPAM_{def}$ を作成する．次に，シソーラスを作成した残りのメールデータであるスパムメール N 個と非スパムメール N 個の合計 2N 個のメールデータ，$SPAM_{est}$，$SPAM_{def}$ から，解析用データを作成する．この解析用データは，スパムメール N 個と非スパムメール N 個の合計である 2N 個作成される．その後，解析用データを 3 分割し，交差検定用データに加工する．3-fold cross validation によって，解析結果の検定を行う．

### 3.1　シソーラスの作成

5 セットのメールデータそれぞれについて，スパムメールデータと非スパムメールデータの差分である M 個のデータをスパムメールデータからランダムに選出する．形態素解析ツールである茶筌 [8] を用いて，選出されたメール本文を単語に分割し，各単語の出現回数をカウントする．スパムメールと非スパムメールの差分である M 個のメールデータから得られた単語を出現回数の降順に並べる．ここから，単語の出現回数により，$SPAM_{est}$，$SPAM_{def}$ の 2 種類のシソーラスを作成する．シソーラスの各単語にはスパム度を設ける．

#### 3.1.1　頻出スパム単語集合 $SPAM_{est}$

$SPAM_{est}$ は，スパムメール本文に頻出する単語をまとめた単語集である．出現回数を降順に並べた単語のうち，全単語数の上位 50%の単語を使用する．この単語の中から，記号，1 文字の平仮名・カタカナ，英数字を取り除き，$SPAM_{est}$ とする．各単語の出現回数を全単語で割った値を，その単語のスパム度として設定する．

#### 3.1.2　確定的スパム単語集合 $SPAM_{def}$

$SPAM_{def}$ は，スパムメールのみに出現すると考えられる単語をまとめた単語集である．出現回数を降順に並べた単語のうち，1 度のみ出現した単語を使用する．この単語の中から，記号，1 文字の平仮名・カタカナ，英数字を取り除く．さらに，非スパムメールにも出現すると考えられる単語は取り除き，$SPAM_{def}$ とする．$SPAM_{def}$ は，スパムメールのみに出現する単語を集めているので，スパム度は 1 とする．つまり，

$SPAM_{def}$ に含まれる単語が現れるメールは，スパムであると判断できる．

### 3.2 解析用データの作成

スパムメール N 個と非スパムメール N 個，$SPAM_{est}$，$SPAM_{def}$ から，解析用データを作成する．解析用データは，スパムメール N 個と非スパムメール N 個の合計である 2N 個である．データセットに入っているメールデータは実際の受信データであるため，それぞれのデータセットの時期によっては受信数にばらつきはある．

解析用データの入力変数を以下に示す[9]．以下の入力変数のうち，IP アドレス・Message-ID・件名・名前・URL の個数・URL の割合・スパム語の出現率・スパムコンテンツ度は数値として扱い，テキストの種類・添付マルチパート・メールの種類は文字列として扱う．

- IP アドレス
  メールデータの Received フィールドから IP アドレスを抽出する．スパムメールは Received フィールドに含まれる全ての IP アドレスを抽出し，非スパムメールは Received フィールドに含まれる IP アドレスのうち，ゲートウェイ以前の IP アドレスを抽出する．抽出した IP アドレスには番号を付け，マッチング表を作成する．
- Message-ID
  メールデータの Message-ID フィールドから抽出する．通常のメールでは，Message-ID フィールドに含まれるドメイン名と送信者のアドレスに含まれるドメイン名が非常に似通ったものになるが，スパムでは Message-ID が受信者のドメイン名や受信者のメールサーバの名称そのものとなることが多い．この特徴を判定材料とするための属性である．この属性値は，Message-ID と送信者のメールアドレスのドメイン部の一致度（数値）とする．
- 件名
  メールデータの Subject フィールドから件名を抽出する．抽出した件名に番号を付け，マッチング表を作成する．スパムメールでは，時期を変えて何度も同じ件名を使っていることが多いので，それを利用するための属性として用意した．
- 名前
  メールデータの From フィールドから名前を抽出する．抽出した名前に番号を付け，マッチング表を作成する．件名と同様，スパムメールでは何度も同じ差出人の名前を使っていることが多いので，その特徴を利用するための属性として用意している．
- テキストの種類
  メールデータの Content-Type フィールドからテキストの種類を抽出する．HTML 形式であれば 1，text 形式であれば，2 とする．スパムメールと非スパムメールでは，テキストの種類が異なる可能性がある．スパムメールと非スパムメールが HTML 形式と text 形式のどちらか一方に振り分けられていれば，重要な判定要素として利用することができる．
- 添付マルチパート
  メールデータの Content-Type フィールドから添付マルチパートを抽出する．添付がなければ 1，text 形式の添付があれば 2，text 形式以外の添付があれば 3 とする．テキストの種類と同様に，スパムメールと非スパムメールでは，添付マルチパートが異なる可能性がある．スパムメールと非スパムメールが添付なし，text 形式の添付，text 形式以外の添付のいずれかに偏って振り分けられていれば，重要な判定要素として利用することができる．
- URL の個数
  本文中に含まれる URL の個数をカウントする．スパムメールでは，本文に URL しか含まれていないものや，全く URL のないものも存在する．
- URL の割合
  本文中に含まれる URL のバイト数を本文のバイト数で割った値である．一般的にスパムメールには，URL が含まれていると言われているが，全く URL がないものもある．一方，非スパムメールであっても，受信者が望んで受信した広告メールなどには，URL が多数存在する．現在のメールと過去のメールでは，本文に含まれる URL の個数に違いがあるとも考えられるので，URL の割合の時期的な傾向や割合の多少によって，スパムメールと非スパムメールの特徴をつかむ．
- スパム語の出現率
  シソーラスに含まれる単語のうち，本文中に含まれている単語のバイト数を本文のバイト数で割った値である．この値が大きいメールは，頻出スパム単語集合 $SPAM_{est}$・確定的スパム単語集合 $SPAM_{def}$ に含まれるスパム語が多く存在するということなので，そのメールがスパムであるということを判定する要素として使用することができる．
- スパムコンテンツ度 $SPAM_{grade}$
  本文中に含まれるシソーラスの単語のスパム度を合計した値である．スパム度には $SPAM_{est}$，$SPAM_{def}$ それぞれに重み $w_{est}$，$w_{def}$ をつけて

次のように計算する．

$$SPAM_{grade} = w_{est} \sum_{i} SPAM_{est}^{i} + w_{def} SPAM_{def} \quad (1)$$

$SPAM_{est}$，$SPAM_{def}$ に重みをつけたのは，$SPAM_{def}$ については，単語のスパム度を 1 としているので，$SPAM_{def}$ の単語が含まれるメールは全てスパムと判定されるが，$SPAM_{est}$ の単語が含まれるメールは，$SPAM_{def}$ に対して $SPAM_{est}$ のスパム度が低くなってしまうからである．重みをつけることによって，$SPAM_{est}$ と $SPAM_{def}$ の単語のスパム度を同等に扱えるようにした．

- メールの種類

  スパムメールであれば spam，非スパムメールであれば non-spam とする．この spam，non-spam の値は，メールの受信者が，スパムメールであるか，非スパムメールであるかを判断して付与した値である．この値を参照することで，解析をした結果，誤ってスパムまたは非スパムと判別されたメールデータを確認することが可能である．

### 3.3 交差検定用データへの加工

作成した解析用データは，データマイニングツールである MUSASHI [10] を用いて交差検定用データに加工し，3-fold cross validation を使って，解析結果（識別率）の検定を行う．作成したそれぞれの解析用データを 3 分割し，これを 1 セットとする．3 分割したデータのうち，2 つを学習データ，残りの 1 つをテストデータとする．この学習データとテストデータの組み合わせは，AB と C，BC と A，CA と B の 3 通りできる．

## 4. 解析と結果

### 4.1 解析

データの解析には，データマイニングツールである Weka [11] を使用する．学習アルゴリズム J48 を用いた解析結果から，学習データとテストデータの識別率，スパムデータ・非スパムデータと判定した根拠となる入力属性の期間変化について検討する．式 (1) 中の $w_{est}$ と $w_{def}$ は 50:1 として解析を行った．

使用するメールデータは，2008 年 12 月から 2010 年 2 月までの 15ヶ月間に受信したもので，その内訳は表 1 のとおりである．この期間において，ほぼ毎月 100 通前後の日本語のスパムメールを受け取っており，受け取ったメールがスパムメールかどうかの判断は著者ら自身によって行っている．従って，使用したスパムメールのデータは，明らかにスパムと判断されるもののみである．

スパムメールはすべて日本語のものとし，非スパムメールも同じ時期のものからピックアップしたものである．

表 1　解析したメールの内訳
Table 1　number of analyzed mail dataset

| 受信時期 | メールの個数 | |
|---|---|---|
| | spam | non spam |
| 2008/12〜2009/02 | 395 | 332 |
| 2009/03〜2009/05 | 333 | 289 |
| 2009/06〜2009/08 | 298 | 253 |
| 2009/09〜2009/11 | 216 | 183 |
| 2009/12〜2010/02 | 647 | 527 |

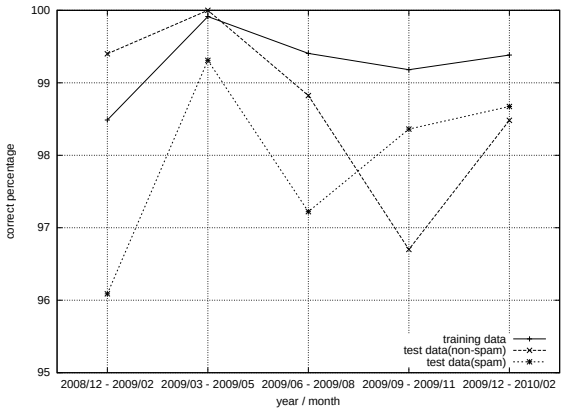


図 2　メールの到着時期による正解率の変化
Fig. 2　correct percentage of term

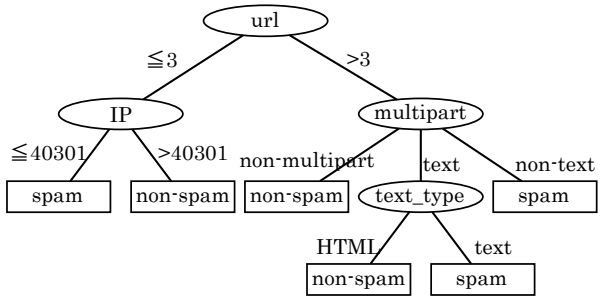


図 3　決定木 1
Fig. 3　decision tree, No.1

### 4.2 結果

学習アルゴリズム J48 の学習データ，テストデータのスパム・非スパムの識別率を図 2 に，2009 年 6 月〜2009 年 8 月，2009 年 9 月〜2009 年 11 月，2009 年 12 月〜2010 年 2 月のメールデータの解析結果から得られた決定木をそれぞれ図 3，図 4，図 5 に，全メールデータに対する text 形式の割合，HTML 形式の割合，添付がないメールの割合，text 形式の添付があるメールの割合をそれぞれ図 6，図 7，図 8，図 9 に示す．

## 5. 考察

### 5.1 メールデータの識別率と決定木の検定

図 2 より，いずれのデータセットについても，スパムデータ，非スパムデータ共に識別率が 96%以上であった．3-fold cross validation を使って解析結果の検定を

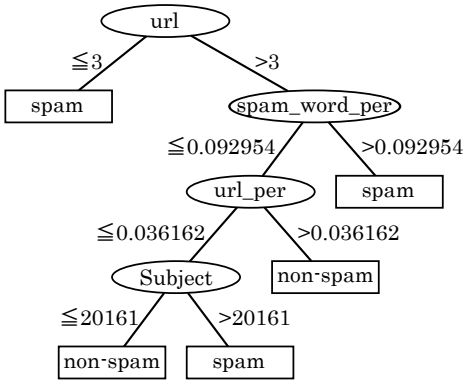


図 4　決定木 2
Fig. 4　decision tree, No.2

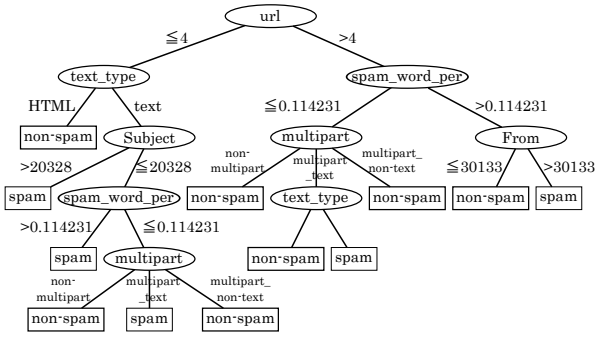


図 5　決定木 3
Fig. 5　decision tree, No.3

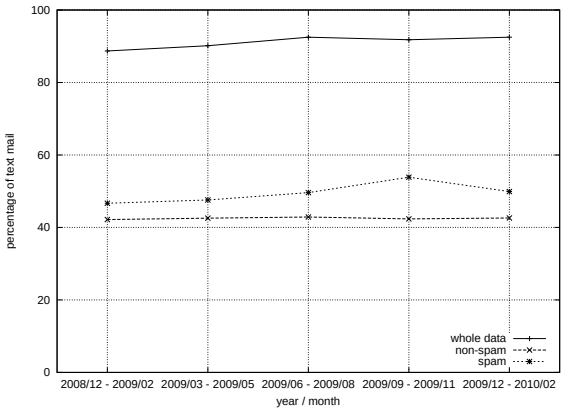


図 6　テキストメールの割合
Fig. 6　percentage of text mail

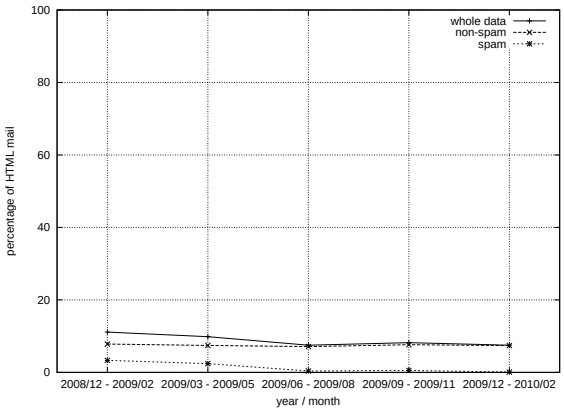


図 7　HTML メールの割合
Fig. 7　percentage of HTML mail

行ったが，いずれの試行においても大きな差異が見られず，同様の識別率を示しており，検定結果としては良好であった．また，モデルの学習に使用したアルゴリズムは J48 であるが，交差検定のいずれの試行においても，同様の決定木を生成しているため，この解析結果は妥当なものであると考えられる．

### 5.2　時間経過に伴なうスパムメールの傾向の変化

それぞれの期間ごとの決定木より，表 2 に示すような判別を行うための主な入力属性が得られた．これによると，2009 年 6 月〜2009 年 8 月（図 3）では，URL の個数が 3 個以内でスパムメールの 91%，4 個以上で非スパムメールの 94%が判別できている．2009 年 9 月〜2009 年 11 月（図 4）では，URL の個数が 3 個以内ではスパムメールの約 8 割が判別できており，残りの 2 割はスパム語の出現率で判別できている．非スパムメールのほとんどは，URL 率によって判別できている．2009 年 12 月〜2010 年 2 月（図 5）については，URL の個数，スパム語の出現率，テキストタイプで分類した後，件名で分類することによりスパムメールを，マルチパートの種類で非スパムメールを判別できている．

図 3〜図 5 に示す決定木は，3ヶ月ごとの連続したメールデータから構築されたものであるが，その期間において徐々に決定木が複雑化していることが分かる．これにより，受信時期が新しくなるにつれて，スパムメールと非スパムメールを判別するためにより多くの入力属性を必要としていることが分かった．また，本論文に掲載していない図 3 以前の期間での決定木は図 3〜5 の決定木の部分木を構成している．従って，本研究で扱った 3ヶ月ごとの期間での解析では，それ以前の決定木，つまり判別モデルを取り入れた決定木を構成していることが明らかになった．

本研究の検証に用いた 15ヶ月のメールデータに関しては，件名が決定木に現れた箇所は 3 箇所しかなく，スパムメールと非スパムメールを直接的に判別する根拠となったものは 1 箇所のみで，スパムメール，非スパムメールともに数個のデータを判別する箇所であった．このことより，件名という属性はスパムメールを判別するための大きな要因とはなっていないが，他の要因と組み合わせることによって判別モデルを構成する必要があるということを示していると考えられる．

### 5.3　テキストの種類と添付マルチパート

図 6，図 7 より，メールデータのテキスト種類は，大部分が text 形式であることがわかった．text 形式におけるスパムメールの割合と非スパムメールの割合は，

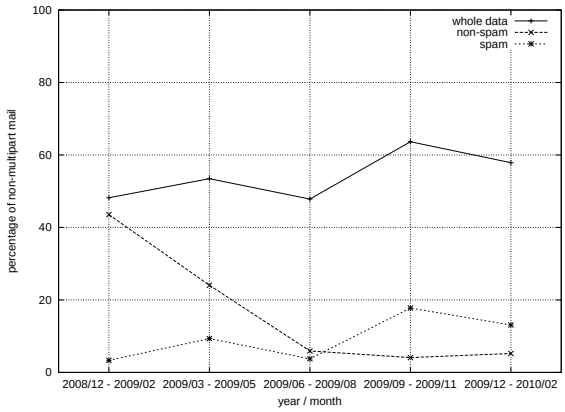


図 8　添付ファイルのないメールの割合
Fig. 8　percentage of non-multipart mail

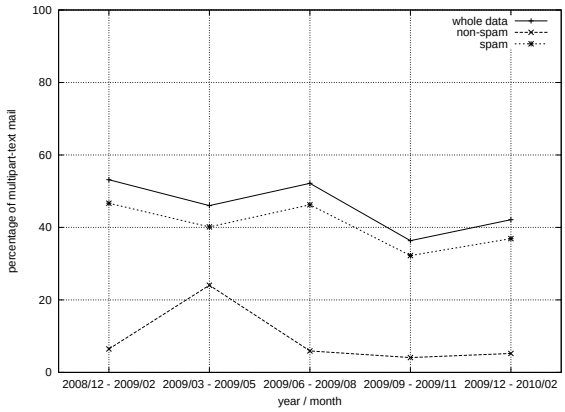


図 9　テキストの添付ファイルのあるメールの割合
Fig. 9　percentage of text multipart mail

表 2　判別に使用した主な属性
Table 2　principal properties used for distinction

| 受信時期 | 入力属性 |
|---|---|
| 2008/12〜2009/02 | IP |
| 2009/03〜2009/05 | マルチパートの種類，名前，IP |
| 2009/06〜2009/08 | URL の個数，IP アドレス |
| 2009/09〜2009/11 | URL の個数，スパム語の出現率，URL 率 |
| 2009/12〜2010/02 | URL の個数，スパム後の出現率，マルチパートの種類 |

ほぼ同じであるが，HTML 形式のものはスパムメールに対して，非スパムメールの方が圧倒的に多い．図 8，図 9 より，メールデータの添付マルチパートは，スパムメールでは text 形式の添付があるものが多く，非スパムメールでは添付がないものが多い．したがって，テキストの種類と添付マルチパートは，スパムメール・非スパムメールを判別するための重要な要素の 1 つであると言える．

## 6.　おわりに

メールデータの判定要素として，スパムデータはテキストの種類が text 形式，添付マルチパートの text 形式の添付があるもの，非スパムメールデータではテキスト種類の HTML 形式，添付マルチパートの添付がないもの，スパム語の出現率が特に重要である．また，URL の個数と URL の割合，スパム語の出現率，スパムコンテンツ度により，ある程度スパムメールと非スパムメールの判定が可能である．今回，学習アルゴリズムは J48 のみを使用して判定要素の検討を行ったが，他の学習アルゴリズムと比較することが必要であると考えている．

また，過去 15ヶ月分のデータを 3ヶ月ごとに解析をしたが，スパムメールの特徴を細かく分析するには，1ヶ月ごとに解析をして，結果を比較することも必要である．IP アドレスの抽出について，スパムメールは Received フィールドに含まれる全ての IP アドレスを抽出し，非スパムメールは Received フィールドに含まれる IP アドレスのうち，ゲートウェイ以前の IP アドレスを抽出したが，非スパムメールにおいて，ゲートウェイ後の IP アドレスを抽出して，同様に解析をすると結果が変化する可能性がある．

## 参考文献

[1] 渡部綾太, 愛甲健二: スパムメールの教科書, データハウス, (2006).
[2] 吉澤亨史: CNET Japan, http://japan.cnet.com/ news/business/story/ 0,3800104746, 20416931,00.htm, 2010 年 7 月 16 日確認.
[3] S. Lee, et.al: Spam Detection Using Feature Selection and Parameters Optimization, CISIS 2010, pp.883-888, (2010).
[4] T. Almeida, Akebo Yamakami, J. Almeida: Filtering Spams using the Minimum Description Length Principle, SAC'10, pp.1854-1858, (2010).
[5] Y. Hu, el.al: A scalable intelligent non-content-based spam-filtering framework, Expert Systems with Applications, Elsevier, (2010).
[6] M. Soranamageswari, C. Meena: Statistical Feature Extraction for Classification of Image Spam Using Artificial Neural Networks, ICMLC, pp.101-105, (2010).
[7] I. Hayashi, T. Kobayashi, Y. Arai, T. Maeda, A. Inoue: Optimal Location of Wireless LAN Access Points Using Fuzzy ID3, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.13, No.2, pp.128-134, (2009).
[8] chasen legacy, http://chasen-legacy.sourceforge.jp/, 2010 年 7 月 16 日確認.
[9] 佐々木梨絵: スパムメールを識別するための判定要素の検討, 2009 年度卒業研究.
[10] MUSASHI, http://musashi.sourceforge.jp/, 2010 年 7 月 16 日確認.
[11] Weka 3, http://www.cs.waikato.ac.nz/ml/weka/, 2010 年 7 月 16 日確認.