O-012

迷惑メールフィルタリングアルゴリズムの評価用メールセット
Mail Sets for Spam Filtering Tests

鈴木 貴史† 白石 善明† 溝渕 昭二‡
Takashi Suzuki Yoshiaki Shiraishi Shoji Mizobuchi

1. まえがき

迷惑メールの手口は巧妙化する一方で，サービスプロバイダなどのサーバに迷惑メール対策の機能があってもそれで対処できないものが新たに出現する可能性がある．サーバで対処できないものに対して，残されている対策方法は利用者の PC で迷惑メールであるかどうか判断するプログラムを実行し，最終的にはメールの受信者が判断することである．

そのような立場から，エンド PC で利用する迷惑メールのフィルタリングアルゴリズムの研究がなされている．しかし，その評価のためのメールセットが異なっていることも多く，アルゴリズムの性能比較を行うことは容易ではない．メールセットにはメールの内容や送受信関係に何らかの特徴が存在することもあり得るので，フィルタリングアルゴリズム間の優劣を比較する際に，使用するメールセットとして個人的に過去に受信したメールを集めたもので評価しないことが望ましい．

そこで，様々な傾向はパラメータによって調整できるようにし，複数のデータセットで各フィルタリングアルゴリズムを評価してアルゴリズムの有効性を検証するための特定の個人に依存しないメールのデータセット生成法を提案する．

2. 従来の評価方法

フィルタリングアルゴリズムの評価実験を行うときに使用されるメールのデータセットは各論文で異なっているので，比較のための再現実験を行うことができない場合が多い．例えば，http://spamlinks.net/filter-archives.htm のように，迷惑メールのサンプルをある特定のプロジェクト・個人が収集したものを提供している場合があるので，このような公開されているものを利用する方法が考えられる．しかし一方で，メールセットに含まれるべき正当なメールのセットを公開しているところはあまり存在しない．一般に公開しているメーリングリストのメールを用いることも考えられるが，個人的なメールのやり取りとは異なる内容になってしまうことが懸念される．

カーネギーメロン大学の Enron Email Dataset を利用した論文 [1] がある．Enron のデータセットは 158 人から収集した 200,399 通をまとめたことにより一般性を持たせようとしている．しかし，このような大規模なデータセットの利用は，どちらかといえばサーバなどで用いられるフィルタリングアルゴリズムの評価に適しており，エンド PC 上で利用するアルゴリズムの傾向を把握できない可能性がある．しかも，あるメールが迷惑メールであるかどうかの判断は自分でしなければならないので，評価の前の事前準備に手間がかかる．

†名古屋工業大学, Nagoya Institute of Technology
‡近畿大学, Kinki University

3. 評価用メールセットの自動生成

3.1 メールデータの構成

評価用メールセットに含まれるデータである電子メールは，そのヘッダと本文を RFC2822 [2] に従って生成する．RFC2822 のうち，しなければならない (MUST) ものと，するべきである (SHOULD) ものに注意し，特に次の規則に沿ってメールデータを生成する．

- 1 行は CRLF を除き 998 文字以下 (78 文字以下推奨) とする
- US-ASCII で記述する
- ヘッダフィールドは “フィールド名:フィールドボディCRLF” で記す
- ヘッダ部とボディ部は空行で区切る

以下では，ヘッダ部とボディ部について述べる．

3.1.1 ヘッダ部

例えば，文献 [5] では，ヘッダ部に含まれる情報からメールの送受信関係に基づくアルゴリズムが提案されている．そこで，From，To，Cc のフィールドの自動生成に対応する必要がある．

送受信関係として図 1 のように 3 つのパターンを定義する．

[パターン 1] 送信者は自分で，受信者は任意の数の他人
[パターン 2] 送信者は他人で，受信者は自分と任意の数の他人
[パターン 3] 送信者は他人で，受信者は自分を含まない任意の数の他人

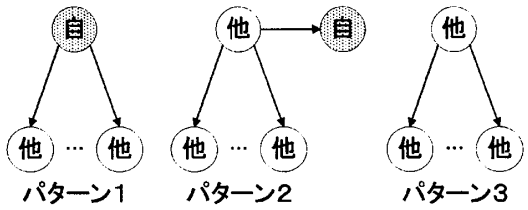

図 1: メール送受信関係の基本パターン

各パターンの各ノードに適当なアドレスをあてはめることにより，一般性を失うことなくメールの送受信関係を表現できる．

3.1.2 ボディ部

文献 [3] 等で，特徴的な単語の出現頻度によるフィルタリングアルゴリズム (ベイジアンフィルタ) が提案されており，そのようなものにボディ部は対応する必要がある．

ボディ部の生成には，まず，ベイジアンフィルタなどで，適当なサンプルメールセットを用いて，あらかじめ正当メールと迷惑メールに出現しやすいそれぞれの単語のデータベースを学習により構築しておく．そして，それぞれのデータベースから多数の単語を選択し，スペースで区切って単語を並べることにより，ボディ部を生成する．以上がボディ部の基本的な生成方法である．

3.2 データセットのパラメータ

パラメータの調整によって作成されるメールセットが変化する．パラメータを変更したメールセットを複数用意し，フィルタリングをしてみることで，そのフィルタリングアルゴリズムの多面的な評価が可能となる．

生成するメールセットのパラメータは総量で指定するものと割合で指定するものに分けている．

[総量で指定するパラメータ]

- N 作成するメールの数
- A_H 正当なメールを送信するアドレスの数
- A_{HM} A_H の内，送信数の多いアドレスの数
- A_S 迷惑メールを送信するアドレスの数
- A_{SM} A_S の内，送信数の多いアドレスの数
- W_H ボディ部に含まれる正当な単語の数
- W_S ボディ部に含まれる迷惑な単語の数

[割合で指定するパラメータ]

- P_R パターン 2 の正当メールに本人が返信する割合
- P_{SHD} 迷惑メールの内，ヘッダが偽造されている割合
- P_{HM} 正当なメールの内，A_{HM} が送信する割合
- P_{SM} 迷惑メールの内，A_{HS} が送信する割合
- P_{O_i} 当該パターンの宛先に含まれる自分以外のアドレスの数 i の割合
- P_{H_n} パターン 1,2,3 の正当メールを生成する割合
- P_{S_n} パターン 2,3 の迷惑メールを生成する割合

3.3 生成手順

メールデータを自動生成するために，各パラメータを適宜読み込む．例えば，ボディ部では正当メールか迷惑メールのどちらを作るかに応じて W_H，W_S を変更し，生成することや，パターン 2 の不当なものを作るときに，ヘッダの偽装により送信者が本人になるものが一定の確率で生成されるようにする等である．

以下のような手順で電子メールを自動生成する．

- Step1 $N \times P_{H_n}$ と $N \times P_{S_n}$ から各パターンの生成数 N_{P_i} を計算
- Step2 $N_{P_i} \times P_{O_i}$ で宛先数別の生成数を計算
- Step3 パターン 2 の正当メールを生成する．同時に返信を作成し，パターン 1 の正当なものを作ったものとして扱う．
- Step4 パターン 1 の正当メールを生成する．
- Step5 パターン 3 の正当メールを生成する．
- Step6 パターン 2 の迷惑メールを生成する．
- Step7 パターン 3 の迷惑メールを生成する．

特に Step3 の内容は Step4 の内容よりも前に実行する必要がある．なぜなら，生成するメールの総数が決まっている場合，パターン 1 の正当なものをパターン 2 の正当なものより先に生成してしまうと，パターン 2 の正当なものに対する返信 (パターン 1 の正当なものに相当する) が生成できなくなってしまうからである．

3.4 サンプル

生成されたメールデータの例を以下に示す．ヘッダフィールドは RFC2822 に規定されているもののうち，Subject，From，To，Cc を生成している．

生成された正当メールの例

```
Subject: This Mail is ham
From: "知り合い"<000001@ham>
To: "受信者本人"<user@nitech.ac.jp>,
   "知り合い"<000031@ham>

linux server please
```

生成された迷惑メールの例

```
Subject: This Mail is spam
From: "スパマー"<000001@spam>
To: "受信者本人"<user@nitech.ac.jp>,
   "知り合い"<000047@ham>

spy waiter posts
```

3.5 ファイル名

フィルタリングをするとき，3.3 節で生成された順でメールデータを入力すると，何らかの性質を有するフィルタリングアルゴリズムを有利あるいは不利にしてしまう可能性がある．そこで，ファイル名の先頭に乱数を用いる．後にプログラムに入力するファイル一覧を作成するときに，文字コード順にソート済のファイル一覧を OS から取得することで，生成順とは全く異なる順番に読み込むことができる．

また，あるメールに対する返信を直後に生成することができたり，乱数が重複した場合に上書きされることを防ぐために整理番号を付ける．

そして，そのメールデータが正当か迷惑かの区別をするために，拡張子の前に ham か spam の文字列を挿入する．

以上をまとめると，ファイル名は次のようになる．

乱数-整理番号-{ham,spam}.eml

4. 生成したメールセットによる比較

4.1 比較対象

生成したメールセットの有効性を確認するために，1) ベイジアンフィルタによる方法 [3]，2) クラスタリング係数による方法 [5]，3) 受信スコアによる方法 [6] の 3 つを利用する．

1) ベイジアンフィルタによる方法

正当なメール (Ham) と迷惑なメール (Spam) を単語に分解し，単語について統計を取ると，ある単語 w について，正当なメールに出現する確率と迷惑メールに出現する確率を求めることができる．ある単語が迷惑メールに出現する確率 $p(w)$ を用いて，受信したメール m が迷惑メールである確率 $p(m)$ を計算する．その確率が閾値 t を越えたものを迷惑メールと判断する方法である．確率 $p(w)$ は学習データとして保存される．

なお，ここでは bsfilter [4] というソフトウェアで，アルゴリズムは Gary Robinson-Fisher 方式を用いる．

2) クラスタリング係数による方法

以下では CCM(Clustering Coefficient-based Method) と呼ぶ．

あるメールが迷惑メールであるかの判断をするために，メールの送受信関係から求められた各ノードのクラスタリング係数を用いる手法 [5] である．クラスタリング係数とはネットワークの緊密度を示す指標である．

送受信関係を表すグラフをクラスタリングし，その各部分ネットワークに含まれる全てのノードのクラスタリング係数を平均する．その値がある値 T_U 以上の場合にはその部分ネットワークに含まれるノードは正当，T_L 未満の場合には迷惑メールを送信するノード，T_U 未満かつ T_L 以上の部分ネットワークであればそれらに含まれるノードは正当でも迷惑でもなく不定とするものである．

3) 受信スコアによる方法

以下では，RSM(Receiving Score-based Method) と呼ぶ．

あるメールが迷惑メールであるかの判断をするために，送受信関係を重み付き有向グラフから得られる各ノードのスコアを用いる手法 [6] である．

求めた受信スコアは，図 2 のようにソートをすると大きく 3 つに分類される．水平直線上に並んでいるスコアをスパム値と呼び，このスパム値を持つノードに迷惑メールの送信者が含まれる．

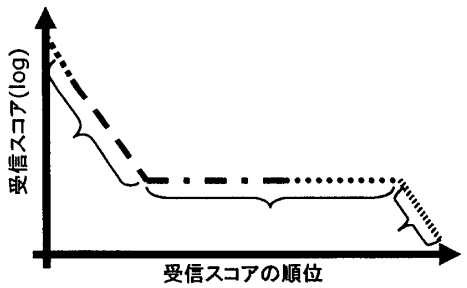

図 2: 受信スコアの分布

上記 3 つの手法をはじめとするフィルタリングアルゴリズムの評価には，表 1 に示す 3 つの指標が主に用いられる．誤遮断率 FPR と誤通過率 FNR から誤判断率 ER が得られる．誤判断率が低いアルゴリズムが優れている．

表 1: フィルタリングの評価指標

FPR	false-positive rate 誤遮断率 正当な電子メールが迷惑メールとして遮断されてしまう確率
FNR	false-negative rate 誤通過率 迷惑メールが正当なメールとしてみなされ，フィルタを通過してしまう確率
ER	error rate 誤判断率 迷惑メールであるかどうかを誤判断してしまう確率

4.2 メールセットを生成するパラメータの設定

以下は，今回実験に使用した 3.2 節に示したパラメータの値である．

N=2000, A_H=100, A_{HM}=10, A_S=750, A_{SM}=100, 正当メールの場合 (W_H=80, W_S=20), 迷惑メールの場合 (W_H=15, W_S=85), P_R=20%, P_{SHD}=1%, P_{HM}=25%, P_{SM}=10%

表 2 に，3.1.1 節の 3 パターンに対するパターン別の宛先数の比率 P_{O_i} を示す．表中の他人ノード数とは，宛先のうち自分以外のアドレスの数のことである．

メールセットのそれぞれを構成する各パターンの正当メール，迷惑メールの比率 P_{H_n}, P_{S_n} を表 3 に示した値とする．

表 2: パターン別宛先数の比率 P_{O_i}

他人ノード数		0	1	2	3	4
パターン 1	正当	0%	80%	10%	6%	4%
パターン 2	正当	40%	30%	15%	15%	0%
	迷惑	0%	100%	0%	0%	0%
パターン 3	正当	10%	60%	30%	0%	0%
	迷惑	30%	70%	0%	0%	0%

表 3: メールセットを構成する各パターンの比率 P_{H_n}, P_{S_n}

メールセット	パターン 1	パターン 2		パターン 3	
	正当	正当	迷惑	正当	迷惑
1	60%	10%	10%	10%	10%
2	10%	60%	10%	10%	10%
3	40%	25%	15%	10%	10%
4	25%	40%	15%	10%	10%
5	25%	10%	40%	10%	15%
6	25%	15%	10%	40%	10%
7	20%	20%	20%	20%	20%

以上のパラメータで生成された 7 種類のメールセットにおける，正当メールと迷惑メールの割合を図 3 に示す．

4.3 メールセットの違いによる結果の変化

各アルゴリズムに対して，7 種類のメールセットを入力し，判定成功率，誤判定率，不明率を調べた．

図 4 は CCM によるフィルタリング結果，図 5 は RSM によるフィルタリング結果，図 6 はベイジアンフィルタによるフィルタリング結果である．

CCM によるフィルタリングは図 4 に示すように，メールセット毎の結果が大きく異なっている．CCM は利用者のアドレスを除いてクラスタリング係数を求めるため，利用者のアドレスからのメールは全て不明と判断されてしまう．パターン 1 の比率が大きいメールセットで特に成功率が低くなる傾向があるということがわかる．

RSM によるフィルタリングは図 5 に示すように，メールセット 5，7 を除き，成功率は 80%弱で一定となっている．RSM で迷惑メールは不明と判断されるので，迷惑メールであるパターン 2，3 の比率が他のメールセッ

トに比べて高いメールセットである 5 と 7 は成功率が低くなったと考えられる．

ベイジアンフィルタによるフィルタリングは図 6 に示すように，パターン 5，7 で他より成功率が下がっている．これはメールセット中の迷惑メールの割合が他より多く，ベイジアンフィルタが W_H=15,W_S=85 というパラメータの設定のために誤通過しやすくなっていたためである．

このように，複数のメールセットで評価を行うことで，フィルタリングアルゴリズムの特徴を見ることができるようになる．

なお，CCM と RSM はヘッダ部を，ベイジアンフィルタはボディ部を用いており，対象が違うということから，図 4，図 5，図 6 の結果が得られたとしても，必ずしもベイジアンフィルタがもっとも優れているとは言えないことに注意する必要がある．

5. まとめ

フィルタリングアルゴリズム間の比較評価を可能にするという課題に対して，本稿では，パラメータの調整によって任意の送受信者数，送受信関係，本文の内容を必要な数だけ含むメールセットの生成方法を提案した．

7 種類のデータセットを用いて 3 つの方法を評価し，成功率，誤判定率，不明率はメールセットによって異なることを示した．このことから，フィルタリングアルゴリズムの評価には，特定の個人やグループが収集したような何らかの特徴を有している可能性があるメールセットを用いるのではなく，一般性を持たせた複数のデータセットを用いて評価をする必要があると言える．提案方法により生成したメールセットにより，フィルタリングアルゴリズムの比較評価が容易になることや，個人が使っているメーラに組み込まれた迷惑メールフィルタの特性を知るために利用するなどの応用などが期待される．

今回はエンド PC におけるフィルタリングアルゴリズムを対象としたが，サーバで行う迷惑メール対策でも利用できるようなデータセットを作成することが今後の課題として挙げられる．

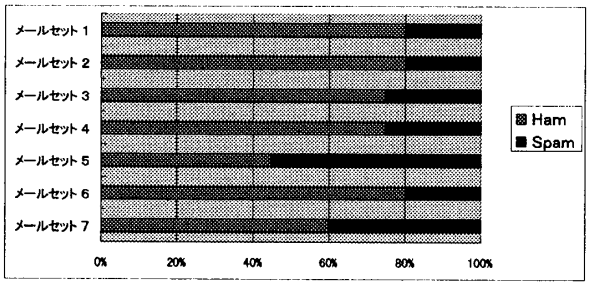

図 3: 各メールセットの正当メールと迷惑メールの割合

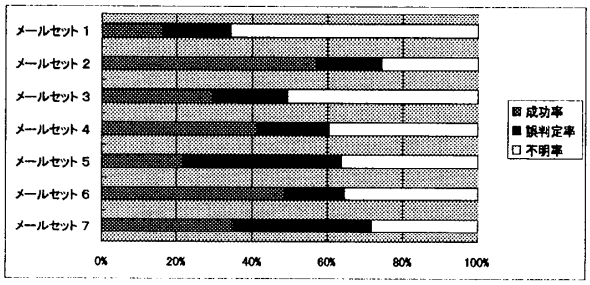

図 4: CCM によるフィルタリングの結果

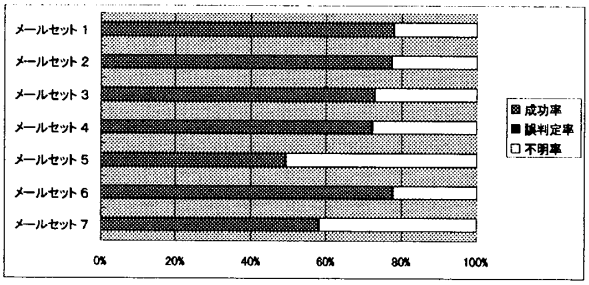

図 5: RSM によるフィルタリングの結果

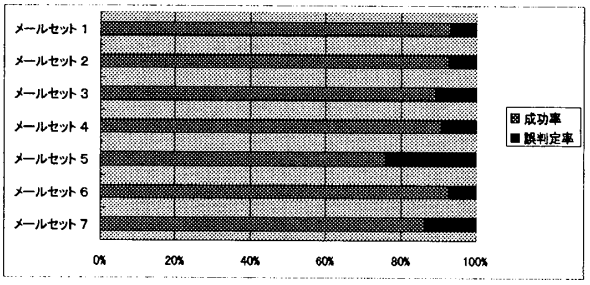

図 6: ベイジアンフィルタによるフィルタリングの結果

参考文献

[1] B. Klimt, Y. Yang, “A paper describing the Enron data,” 2004 CEAS conference, 2004. http://www.ceas.cc/papers-2004/168.pdf

[2] P. Resnick,“Internet Message Format,” RFC2822, April, 2001.

[3] Graham, “A Plan for Spam,” http://paulgraham.com/spam.html

[4] bsfilter / bayesian spam filter, http://bsfilter.org/

[5] P. Oscar Boykin, Vwani P. Roychowdhury, “Leveraging Social Networks to Fight Spam, ” IEEE Computer Society, April 2005.

[6] 鈴木貴史，白石善明，溝渕昭二，“複数フィルタの連携による迷惑メール対策に関する提案と評価，” DICOMO2007，2007.