情報処理学会論文誌　Vol. 52　No. 9　2686–2696 (Sep. 2011)

判定に利用するトークンの限定による ベイジアンフィルタの判定精度向上手法

山内利宏[†1]　上村昌裕[†1]　畑中良太[†2]

迷惑メール対策の１つとしてベイジアンフィルタが利用されている．しかし，迷惑メール送信者は，巧妙な方法で迷惑メールフィルタを回避しようとするため，見逃しメールが発生する可能性がある．そこで，ベイジアンフィルタについて，判定精度の向上を目的として，判定した各メールに含まれるトークンの迷惑メール確率の分布を調査した．調査結果から，見逃しメールの原因の１つとして，初めて出現するトークンの扱いに問題があることを明らかにした．そこで，本論文では，この調査結果を基に，誤検出を増やさずに見逃しメールを減らすことができる判定に利用するトークンを限定した迷惑メール対策を提案する．提案方式は，見逃しメールのトークンの特徴を考慮し，電子メールの迷惑メール確率計算に利用するトークンを限定することで，判定精度を向上させる．複数のメールセットによる評価により，提案方式を用いることで誤検出メールを増やすことなく，見逃しメールを減少させることができることを示す．

Limiting Use of Tokens for Improvement of Bayesian Filter

Toshihiro Yamauchi,[†1] Masahiro Uemura[†1] and Ryota Hatanaka[†2]

Using the Bayesian filter is a popular approach to distinguish between spam and legitimate e-mails. Spam senders sometimes modify emails to bypass the Bayesian filter. The tokens included in the e-mail are investigated for improving the accuracy of classification of emails. The results show that tokens found at the first time sometimes degrade the accuracy of the classification. In this paper, we propose an anti-spam method that consider the difference of the property of tokens. The proposed method limits the use of tokens for improvement of Bayesian filter. The evaluations were performed by using some email sets. The results shows that the proposed method can decrease the false negative rate.

1. はじめに

迷惑メールは，大きな社会問題となっており，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．2010 年 10 月には，電子メール全体の 86.61%を迷惑メールが占めており，内容も多岐にわたる[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，フィッシングメールと呼ばれる詐欺メールも多い[2)]．これらの問題から，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，過去に受信した電子メールから，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）を基に，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測する方式である．ベイジアンフィルタは，フィルタリング精度が高く，個人用途での迷惑メール対策で特に広く利用されている．

しかし，迷惑メール送信者は，ワードサラダ[6),7)] など迷惑メールの内容とは関係のない文章をメールに載せるなど，巧妙な方法で迷惑メールフィルタを回避しようとする．このため，見逃しメール（迷惑メールであるが，誤って正当な電子メールと判定されたメール）が発生する可能性がある[8),9)]．

そこで，ベイジアンフィルタでよく用いられている Robinson-Fisher 方式について，特に見逃しメールに着目し，電子メールに含まれるトークンの特徴を調査した．見逃しメールは，他の電子メールに比べ，初めて現れるトークンを多く含むという特徴があり，これが見逃しメールの原因となることを明らかにした．本論文では，この調査結果を基に，誤検出を増やさずに見逃しメールを減らすことができる，判定に利用するトークンを限定した迷惑メール対策を提案する．提案方式は，見逃しメールのトークンの特徴を考慮し，電子メールの迷惑メール確率計算に利用するトークンを限定することで，判定精度を向上させる．また，異なる方法で収集した複数の電子メール群に対しても評価を行い，提案方式により判定

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

精度が向上することを評価により示した．

2. ベイジアンフィルタ

2.1 概　　要

ベイジアンフィルタは，過去に受信した正当な電子メールと迷惑メールのテキストデータを基に，新たに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測する手法である．ベイジアンフィルタの処理は学習処理と判定処理に分かれており，学習処理では，過去に受信した正当な電子メールと迷惑メールを基に，トークンの迷惑メール確率を格納するコーパスを作成する．判定処理では，作成したコーパスを基に，新たに受信する電子メールの迷惑メール確率を計算する．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[3]，Robinson 方式[4]，および Robinson-Fisher 方式[5] が多く用いられている．これらの計算方式において，各確率は 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，迷惑である可能性が高いことを意味する．

2.2 Robinson-Fisher 方式

本研究で対象とした Robinson-Fisher 方式について説明する．Robinson-Fisher 方式は，トークンの迷惑メール確率 $f(w)$ 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 度も電子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 が出現した回数とする．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．

トークンの迷惑メール確率の計算方法においては，出現回数の少ないトークンの扱いが課題となる．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，トークンの迷惑メール確率 $p(w)$ が 1 となる．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．

一方，Robinson 方式と Robinson-Fisher 方式では，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．このため，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 $p(w)$ の値に近づく．また，トークン w の出現回数が 0 の場合，トークンの迷惑メール確率は x となる．

Robinson-Fisher 方式の電子メールの迷惑メール確率は次の I で与えられる．

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

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

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

C^{-1} は逆 χ^2 関数（inverse chi-square function）を意味する．S は Spamminess（スパム性），H は Hamminess（ノンスパム性）の略で，I はそれらを統合した指標（Indicator）である．

Graham 方式では，特徴的な 15 個のトークンを利用し，それらの結合確率をとることにより，電子メールの迷惑メール確率を計算する．特徴的なトークンとは，トークンの迷惑メール確率が 0.5 から遠く離れているトークンを示す．一方，Robinson 方式と Robinson-Fisher 方式では，すべてのトークンを利用し，式 (3)〜(5) で電子メールの迷惑メール確率を計算する．

3. トークンの迷惑メール確率の調査

3.1 目　　的

メールの迷惑メールの計算は，トークンの迷惑メール確率を基に計算されるため，トークンの迷惑メール確率が判定結果に大きな影響を与えている．誤検出メール（正当な電子

表 1　デフォルトで計算した判定メールの迷惑メール確率の分布
Table 1　Distributions of spam probability calculated by default method.

迷惑メール確率	0.0 以上 ～0.1 未満	0.1 以上 ～0.2 未満	0.2 以上 ～0.3 未満	0.3 以上 ～0.4 未満	0.4 以上 ～0.5 未満	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
正当な電子メール	187 (98.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.5%)	2 (1.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
迷惑メール	9 (0.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	87 (2.1%)	20 (0.5%)	21 (0.5%)	25 (0.6%)	486 (11.9%)	3,435 (84.1%)

メールであるが，誤って迷惑メールと判定されたメール）は正当と判定されたメールに比べ，トークンの迷惑メール確率が高いトークンを多く含み，見逃しメールは迷惑と判定されたメールに比べ，トークンの迷惑メール確率が低いトークンを多く含むと考えられる．そこで，各電子メールに含まれるトークンの迷惑メール確率を調査し，特徴を分析した．

3.2　トークンの分類

電子メールに含まれるトークンを，トークンの迷惑メール確率 $f(w)$ と学習回数を基に以下の 4 種類に分類した．

(1)　正当な電子メールに多く現れるトークン
トークンの迷惑メール確率が 0.5 より小さな値をとるトークンを指す．正当な電子メールに現れる回数が多くなればなるほど，値は 0 に近づく．

(2)　迷惑メールに多く現れるトークン
トークンの迷惑メール確率が 0.5 より大きな値をとるトークンを指す．迷惑メールに現れる回数が多くなればなるほど，値は 1 に近づく．

(3)　両方に同程度現れるトークン
トークンの迷惑メール確率が 0.5 に近い値をとるトークンを指す．正当な電子メールと迷惑メールに現れる割合が同程度の場合，トークンの迷惑メール確率は 0.5 程度となる．

(4)　初めて現れるトークン
コーパスに学習されていないトークンを指す．設定された初期値をとる．初期値は正当と判定されるように偏りを持たせる場合が多く，設定例として，0.4 や 0.415 がある．

本研究で対象とする Robinson-Fisher 方式では，トークンの迷惑メール回数や出現回数を考慮した処理を行っているものの，迷惑メール送信者はフィルタを回避するように工夫して送信してくるため，見逃しメールが発生する．そこで，トークンの迷惑メール確率に着目し，正当な電子メールあるいは迷惑メールによく現れるトークンの特徴を調査した．

3.3　デフォルト設定の判定精度調査

ベイジアンフィルタを採用している bsfilter[10] のデフォルト設定（以降，デフォルトと略す）で調査を行った．デフォルトでは，Robinson-Fisher 方式を採用し，トークンの迷惑メール確率が 0.4 未満 0.6 以上のトークンを判定に利用する．トークンの抽出には，bsfilter のデフォルトで用いる bigram を用いた．bigram は，日本語については，孤立した漢字および 2 字が連続する漢字，連続するカタカナはそのまま 1 つのトークンとして抽出する．また，3 字以上の連続する漢字については 1 文字目と 2 文字目，2 文字目と 3 文字目というように隣接する 2 字の漢字をそれぞれ 1 つのトークンとして抽出する．英単語については空白などで区切られた 1 単語を 1 つのトークンとして抽出する．bsfilter では，日本語の文字コードの電子メールとそれ以外の文字コードの電子メールについて分けてトークンを学習し，コーパスを作成する．

実験に用いた電子メールは，著者らが受信した正当な電子メールと迷惑メールである．2008 年 4 月～8 月に受信した正当な電子メール 1,049 通（日本語 1,006 通，非日本語 43 通）と 2008 年 8 月に受信した迷惑メール 4,687 通（日本語 1,270 通，非日本語 3,417 通）を学習させ，2008 年 9 月に受信した正当な電子メール 190 通（日本語 184 通，非日本語 6 通）と迷惑メール 4,083 通（日本語 492 通，非日本語 3,591 通）を判定させた．計算した判定メールの迷惑メール確率の分布を**表 1** に示す．閾値を 0.9 とした場合，誤検出メールは発生せず，見逃しメールは 162 通（4.0%）となった．

3.4　累積度数分布調査

調査で発生した見逃しの要因を分析するため，トークンの迷惑メール確率の累積度数分布を調査した．この調査では，正当な電子メール，誤検出メール，見逃しメール，検出スパム（正しく検出できた迷惑メール），全迷惑メール（見逃しメール＋検出スパム）に含まれるトークンの違いを明らかにすることを目的としている．3.3 節の調査では誤検出メールがな

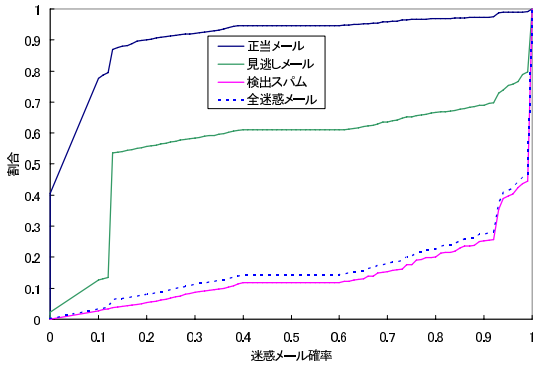


図 1　トークンの迷惑メール確率の累積度数分布
Fig. 1　Cumulative frequency distribution of spam probability of tokens.

かったため，誤検出メールについては調査していない．**図 1** に判定メールに含まれるトークンの迷惑メール確率の累積度数分布を示す．累積度数分布の図は，横軸の迷惑メール確率以下のトークンの総数が，全体のどのくらいの割合を占めるのかを縦軸の値で表している．

図 1 より，正当な電子メールは，トークンの迷惑メール確率が 0.2 未満のトークンが全体の約 90%，0.5 未満のトークンが約 95%を占めることが分かる．正当な電子メールらしさの高いトークンが多いため，電子メールの迷惑メール確率が低くなると考えられる．さらに，トークンの迷惑メール確率がきわめて低い 0.00001 未満のトークンが全体の約 40%を占めている．一方，トークンの迷惑メール確率が高い 0.9 以上のトークンは約 2.7%しか含まない．

全迷惑メールは，トークンの迷惑メール確率が 0.9 以上のトークンが全体の約 47%を占めている．0.5 未満のトークンは約 14%であり，約 86%が迷惑メールらしさの高いトークンであるため，電子メールの迷惑メール確率が高くなると考えられる．また，トークンの迷惑メール確率が低い 0.1 未満のトークンは 3.3%しか含まない．

全迷惑メールを見逃しメール，検出スパムごとに見ると，検出スパムは，全迷惑メールの約 96%を占めているため，全迷惑メールとほぼ同じ分布となっている．見逃しメールは，トークンの迷惑メール確率が 0.12 以上 0.13 未満のトークンが全体の約 40%を占める．迷惑メール確率が 0.13 未満のトークンが全体の約 54%，0.5 未満で約 61%を占め，半数以上が正当らしさの高いトークンであるため，電子メールの迷惑メール確率が低くなると考えられる．今回の判定で発生した見逃しメールは，迷惑メール確率が 0.12 以上 0.13 未満のトークンが主な原因となっていると推察できる．

3.5 考　　察

迷惑メール確率が 0.12 以上 0.13 未満となるトークンは，正当な電子メールと迷惑メールの出現比率がおおよそ 8:1 となるトークン，あるいは初めて現れる（コーパスに学習されていない）トークンであった．今回学習で作成したコーパスでは，初めて出現するトークンの迷惑メール確率は，0.12 以上 0.13 未満の範囲の確率であった．

bsfilter は，コーパス作成時，学習されたトークンの迷惑メール確率と初めて出現するトークンの迷惑メール確率を計算する．初めて出現するトークンの迷惑メール確率 $robx$ は，式 (6) で計算する．$robx$ は，1 回だけ学習されたトークンの学習データを基に，初めて現れる（1 回目に現れる）トークンの迷惑メール確率を予測した確率となる．なお，$robx$ は bsfilter で用いられている変数名であり，式 (2) の x に相当するものである．

$$robx = \frac{\sum p(once)}{sum\_once} \tag{6}$$

- $p(once)$：1 回だけ学習されたトークンの迷惑メール確率
- sum_once：1 回だけ学習されたトークンの総数

正当であると 1 回学習されたトークンの迷惑メール確率は 0，迷惑であると 1 回学習されたトークンの迷惑メール確率は 1 の値をとる．式 (6) では，正当であると 1 回だけ学習されたトークンの数が，迷惑であると 1 回だけ学習されたトークンの数よりも多ければ，$robx$ は 0.5 未満の値をとる．反対に，迷惑であると 1 回だけ学習されたトークンの数の方が多ければ，$robx$ は 0.5 より大きな値をとる．$robx$ は言語別に設定され，3.3 節の実験では，初めて出現する日本語のトークンの迷惑メール確率が 0.128875，外国語のトークンでは 0.921186 と計算され，コーパスに保存されていた．

著者らの判定メールでは，見逃しメールに含まれる迷惑メール確率が 0.12 以上 0.13 未満のトークンは，初めて現れるトークンが約 99%を占めていた．つまり，デフォルトで発生した見逃しメールは，初めて現れるトークンを多く含むこと，および初めて現れるトークンの迷惑メール確率が低すぎることが原因となっている．

正しく判定された正当なメールや迷惑メールには，出現回数が多く，はっきりと正当か迷惑の特徴を示すトークンが多く含まれていたため，正しく判定できる．一方，見逃しメールの分析で明らかになったのは，見逃しが起きてしまうメールには，初めて出現するトークンが比較的多く含まれており，その迷惑メール確率が低い場合，多くの見逃しメールを発生させることである．迷惑メール送信者は，ワードサラダなどを挿入したり，単語を改変したり

するなどして，フィルタリングを回避しようとする．改変された単語は，過去に学習されていない可能性が高いので，初めて出現するトークンとして判定され，見逃しの原因になりうると推察する．

4. 提案方式

4.1 概　　要

3 章では，著者らが収集したメールにおける各メールに含まれている特徴を明らかにし，特に見逃しの原因を明らかにした．bsfilter では，トークンの迷惑メール確率が 0.4 未満または 0.6 以上のトークンを利用する．これは，特徴が曖昧な 0.5 付近のトークンを判定に利用しないために決められたと推察できる．このため，判定に利用するトークンの範囲を適切に設定することで，誤検出を増やさずに見逃しを減らすことが可能であると考える．

本論文では，見逃しメールから判定に利用するトークンの迷惑メール確率を算出し，迷惑メールの判定精度を向上させる手法について述べる．

4.2 予備実験

4.2.1 内　　容

利用するトークンを限定することで，判定精度が向上するのかを検証するために，3.3 節の調査環境で，判定に利用するトークンを以下の 4 通りにして判定結果を実験した．学習と判定に用いた電子メールは，3.3 節と同じである．

(A)　0.00001 未満 0.6 以上
(B)　0.10 未満 0.6 以上
(C)　0.12 未満 0.6 以上
(D)　0.13 未満 0.6 以上

利用する正当らしさの高いトークンの範囲は，見逃しメールの原因となった 0.12 以上 0.13 未満のトークンを境界に，判定に利用する範囲をより限定するように決定した．また，検出スパムを減らさず，見逃しメールを減らすことも目的としているため，0.6 以上のトークンを常に判定に利用することとした．

4.2.2 実験結果

3.3 節と同じく閾値を 0.9 とした場合の提案方式とデフォルトの誤検出メールと見逃しメールの数を**表 2** に示す．

(A) の場合，見逃しメールを大幅に減少させている．しかし，誤検出メールが発生しているため，利用すべき方式ではない．(B) と (C) の場合，誤検出メールを発生させず，見逃しメールを大幅に減少させている．これらの方式は，トークンの迷惑メール確率の違いをうまく利用した方式となっている．(D) の場合，デフォルトとほぼ同じ結果となっており，初めて現れるトークン（迷惑メール確率が 0.12 以上 0.13 未満のトークン）を利用することが見逃しメールの原因であるといえる．

表 2　デフォルトと各提案方式における誤検出メールと見逃しメール
Table 2　False positives and false negatives by default and each proposed method.

	デフォルト	(A)	(B)	(C)	(D)
誤検出メール	0	2	0	0	0
見逃しメール	162	22	86	87	141

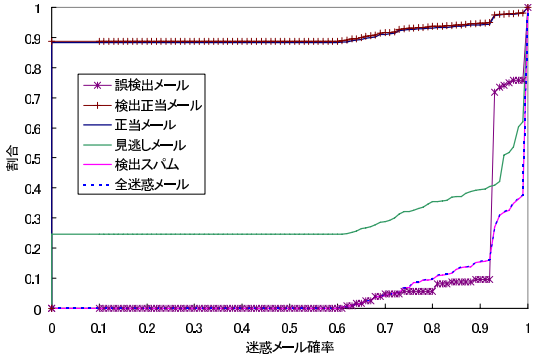


図 2　提案方式におけるトークンの迷惑メール確率の累積度数分布（A）
Fig. 2　Cumulative frequency distribution of spam probability of token by proposed method (A).

4.2.3 各方式におけるトークンの迷惑メール確率の累積度数分布

提案方式で用いたトークンの迷惑メール確率の累積度数分布を (A) は**図 2**，(C) は**図 3**，(D) は**図 4** に示す．(B) は (C) と同じ傾向のため，省略した．なお，(A) は誤検出メールが発生するため，図 2 では，検出正当メール（正しく検出できた正当な電子メール）188 通と誤検出メール 2 通，および正当な電子メール 190 通に関して調査している．

図 2〜図 4 より，利用するトークンを限定するにつれて，見逃しメールは，迷惑メール確率が低いトークンの割合が小さくなることが分かる．したがって，見逃しメールが減少すると考えられる．

(A) が誤検出メールを発生させている原因は，利用するトークンを非常に限定していることによって，ほとんどが迷惑らしさの高いトークンとなっているからである．(B) と (C) の

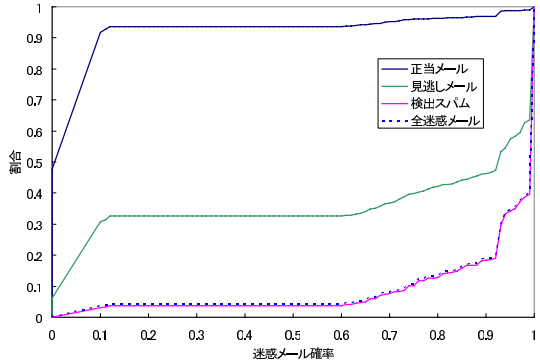


図 3　提案方式におけるトークンの迷惑メール確率の累積度数分布（C）
Fig. 3　Cumulative frequency distribution of spam probability of token by proposed method (C).

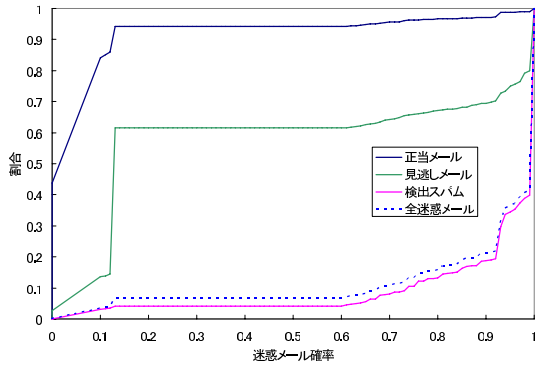


図 4　提案方式におけるトークンの迷惑メール確率の累積度数分布（D）
Fig. 4　Cumulative frequency distribution of spam probability of token by proposed method (D).

場合，見逃しメールの原因となっている初めて現れるトークンを判定に利用していない．このため，迷惑メール確率が低いトークンの割合が小さくなり，見逃しメールが減少したと考えられる．(D) の場合，デフォルトの分布とほぼ同じであるため，見逃しメールもほぼ同じ数になっている．

以上のことから，誤検出を増加させず，見逃しを減らすには，初めて出現するトークンを除外して判定した方が良いといえる．なお，今回の実験では，見逃しの大きな要因が初めて出現するトークンの迷惑メール確率が低く，かつ見逃しメールに含まれる初めて出現するトークンの割合が多いため，このようなことがいえた．したがって，この考え方で判定に利用するトークンを限定する場合，見逃しメールに含まれるトークンの割合を考慮して，判定に利用するトークンを限定するか否か，限定する範囲をどの範囲にするのかを決定する必要がある．以降では，この決定方式について述べる．

4.2.4　異なる電子メールでの実験

さらに，異なる電子メールでの判定精度を調査した．文献 11) で利用されている電子メールを用い，異なる受信環境における判定精度も調査した．文献 11) の電子メールは，文献 11) の筆者らが日常研究活動に用いているメールアドレスに受信した正当な電子メールおよび迷惑メール，および文献 11) の筆者らのうち 1 名が所有するハニーポットアドレスで受信した迷惑メールである．また，非日本語の電子メールとして英語の電子メールを使用している．

文献 11) の正当な電子メール 2,754 通（日本語 1,679 通，非日本語 1,075 通），迷惑メール 1,249 通（日本語 267 通，非日本語 994 通）のうち，半数を学習させ，残りの半数を判定した．デフォルトと比較する提案方式は，4.2.2 項で最も精度が高かった 0.1 未満 0.6 以上のトークンを利用する方式とした．閾値を 0.9 とした場合の正当な電子メールの判定結果は，デフォルトと提案方式で同じであった．また，見逃しメール数はデフォルトで 76 通，提案方式は 59 通であり，見逃しメールを 17 通（約 22%）減少させている．

4.3　判定に利用するトークンを限定する方式

これまでの実験から，判定に利用するトークンを限定することで見逃しメールを減らすことができることが分かったものの，どのような場合に利用するトークンを限定すべきか，限定するとすればどのようにすべきかが課題となる．この課題へ対処した方式について述べる．

あらかじめ正当なメールと迷惑メールを学習させたコーパスが生成されているものとする．

(1)　このコーパスを用いて，bsfilter（デフォルト）で見逃したメールを収集する．
(2)　収集した見逃しメールを bsfilter（デフォルト）で判定する．この際に，判定に利用した各トークンの迷惑メール確率を収集する．
(3)　収集したトークンの迷惑メール確率から，0.01 刻みで迷惑メール確率ごとのトークンの累積度数分布を作成する（図 1 のグラフと同様のデータを作成）．
(4)　累積度数分布の増加量が最も大きい区間を算出する．
(5)　(4) で算出した区間が以下の 2 つの条件をともに満たす場合，限定方式を適用する．それ以外の場合は限定方式を適用しない．

（条件 1） 算出した区間が 0.1 以上 0.4 未満の間に存在する．

（条件 2） 算出した区間の初めて出現するトークンが，判定に利用したトークン全体の 3%以上である．

限定方式を適用する場合，0 以上で算出した区間の下限値未満，および 0.6 以上の迷惑メール確率を持つトークンを，判定に利用するように bsfilter を設定する．限定方式を適用しない場合，デフォルトを用いる．

上記のように，増加量が最も大きい区間を含む範囲の下限を 0.1 としたのは，判定に利用するトークンを限定しすぎると，誤検出が増える可能性があるためである．次に，初めて出現するトークンの割合の下限を 3%としたのは，実験では 5%程度含まれている場合でも十分に効果が確認できており，3%でもある程度の効果が見込まれると判断したためである．また，初めて出現するトークンの割合が少ないにもかかわらず，判定に利用するトークンを限定しすぎると，判定精度が低下する可能性があると考えたためである．上記の手順を自動化するプログラムを設計し，見逃しメール群をこのプログラムで処理することにより，自動的に判定に利用するトークンの範囲を設定することを実現した．

設定の自動化は，上記の 5 つの処理で行われる．処理 (4) と (5) は，C 言語で作成した 46 行のプログラムにより実現した．また，処理 (2)〜(5) を自動化するスクリプト（シェルスクリプトで 6 行）を作成した．処理 (1) では，受け取った迷惑メールの迷惑メール確率を基に，見逃しメールを抽出する．処理 (1) で収集した見逃しメールを用いて，このスクリプトを動作させることで，設定値を算出できる．

5. 評　　価

5.1 評 価 方 法

提案方式の評価を 2 つのメールセットを用いて行った．1 つは，3.3 節で利用したコーパスを用い，著者らが 2008 年 10 月に受信した正当な電子メール 174 通と迷惑メール 5,250 通，および同年 11 月に受信した正当な電子メール 140 通と迷惑メール 2,697 通を判定した．利用するトークンの限定範囲の決定には，2008 年 9 月の見逃しメールを利用した．

もう 1 つは，TREC2007 データセット[12] を用いた．TREC2007 データセットとは，ウォータールー大学の研究グループ宛に 2007 年 4 月 8 日から 2007 年 7 月 6 日までの間に届いた英語の公開メールセットである．

最初に，TREC2007 の 4 月メール 25,622 通（正当メール 6,440 通，迷惑メール 19,182 通）を学習させた後，5 月メール 22,626 通（正当メール 8,710 通，迷惑メール 13,916 通）を判定した．その結果生じた見逃しメールにより，利用するトークンの範囲を変更した後に，5 月メール，6 月メール 22,499 通（正当メール 8,711 通，迷惑メール 13,788 通），および 7 月メール 4,672 通（正当メール 1,359 通，迷惑メール 3,313 通）を判定して，評価した．

次に，さらにメールを受信して学習し，判定に利用するトークンを計算した場合を想定して，TREC2007 の 4 月メールと 5 月メール計 48,248 通を学習させた後，6 月メール 22,499 通を判定した．その結果生じた見逃しメールにより，利用するトークンの範囲を変更した後に，6 月と 7 月メールを判定させて，評価した．

表 3　著者らが受信したメールでの評価結果
Table 3　Evaluation results of emails that authors received.

	デフォルト		提案方式	
	10 月	11 月	10 月	11 月
誤検出メール	0	0	0	0
見逃しメール	500	542	292	330

表 4　TREC2007 での評価結果
Table 4　Evaluation results of TREC2007.

	デフォルト			提案方式		
	5 月	6 月	7 月	5 月	6 月	7 月
誤検出メール	14	10	9	14	10	9
見逃しメール	4,566	6,613	2,103	4,267	6,107	2,053

5.2 著者らが受信したメールでの評価

利用するトークンの範囲を求めた結果，0.12 未満 0.6 以上の迷惑メール確率を持つトークンを利用して，判定した．評価結果を**表 3** に示す．

表 3 から，提案方式では，10 月で 208 通（約 42%），11 月で 212 通（約 39%）の見逃しメールを減少させており，効果が大きいことが分かる．この評価での初めて出現するトークンの迷惑メール確率は，0.128875 であり，除外する範囲に含まれていることが分かった．このことから，見逃しの要因が，初めて出現するトークンの割合が多い（10 月：4.2%，11 月：6.9%）ことによると分かる．

5.3 TREC2007 での評価

TREC2007 の最初の評価において利用するトークンの範囲を求めた結果，0.33 未満 0.6 以上の迷惑メール確率を持つトークンを利用して，判定した．評価結果を**表 4** に示す．

表 4 から，提案方式では，5 月で 299 通（6.5%），6 月で 506 通（7.7%），7 月 50 通（2.4%）

表 5　TREC2007 での評価結果（2）
Table 5　Evaluation results of TREC2007 (2).

	デフォルト		提案方式	
	6 月	7 月	6 月	7 月
誤検出メール	8	7	8	7
見逃しメール	4,380	1,530	4,380	1,530

の見逃しメールを減少させており，一定の効果があることが分かる．この評価での初めて出現するトークンの迷惑メール確率は，0.338777 であり，除外する範囲に含まれていることが分かった．このことから，見逃しの要因が，初めて出現するトークンの迷惑メール確率が低く，その割合が多い（5 月：17.3%，6 月：23.6%，7 月：22.5%）ことであることが分かる．一方，著者らのメールでの評価に比べて効果が小さいのは，初めて出現するトークンの迷惑メール確率が比較的高いためである．

次に，**表 5** に TREC2007 を用いたもう 1 つの場合の結果を示す．この場合，利用するトークンの範囲を求めたところ，増加量が最も多い区間は，0.1 以上 0.4 未満のトークンの迷惑メール確率の範囲に存在しなかった．このため，提案方式でもデフォルトと同じトークンを利用する結果となった．初めて出現するトークンの迷惑メール確率を調べたところ，その確率は 0.555606 となっており，初めて出現するトークンが迷惑メール確率を正当なメールと誤判定させる影響がないことが分かった．この場合は，提案方式を用いてもデフォルトと同じ判定精度である．

図 5 と**図 6** に TREC2007 の評価におけるトークンの迷惑メール確率の累積度数分布を示す．図 5 から分かるように，迷惑メール確率が 0.33 以上 0.34 未満の範囲に多くのトークンが存在しており，提案方式の適用結果から，このトークンが見逃しメールの発生に影響を与えていることが分かる．一方，図 6 では，迷惑メール確率が 0.4 未満の区間において，トークンが集中して分布している箇所はないため，提案方式を適用できず，見逃しメールを減少させることができなかったことが分かる．

5.4　文献 11) のメールでの評価

4.2.4 項の予備実験で利用した文献 11) のメールに対して，提案方式を適用した場合の結果について述べる．デフォルトでは，誤検出メール 5 通，見逃しメールが 76 通であった．提案方式を適用した場合，迷惑メール確率が 0.15 未満および 0.6 以上のトークンを利用して判定した．提案方式では，誤検出メールが 5 通と変わらず，見逃しメールが 67 通となり，9 通（11.8%）削減できていることを確認した．

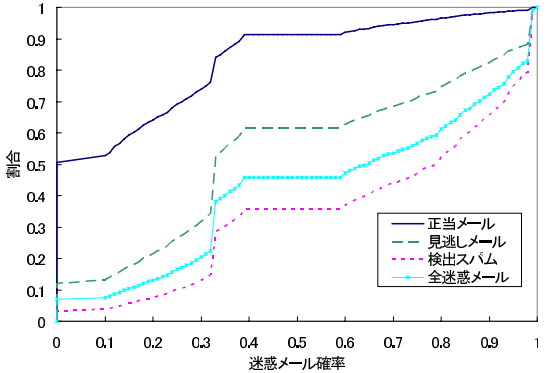


図 5　TREC2007 の評価におけるトークンの迷惑メール確率の累積度数分布
Fig. 5　Cumulative frequency distribution of spam probability of token by TREC2007.

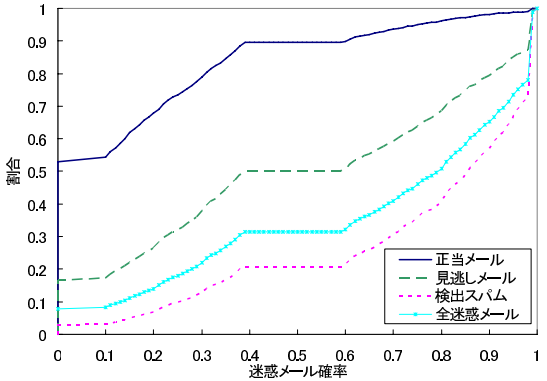


図 6　TREC2007 の評価（2）におけるトークンの迷惑メール確率の累積度数分布
Fig. 6　Cumulative frequency distribution of spam probability of token by TREC2007 (2).

5.5　設定自動化プログラムの処理時間の評価

著者らが受信したメールでの評価（以降，評価 1 と略す），TREC2007 での 2 つの評価（以降，評価 2（5 月，6 月，および 7 月を判定）と評価 3（6 月と 7 月を判定）と略す），および 5.4 節で述べた文献 11) の評価について，作成した設定自動化プログラムの処理時間を評価した．評価は，CPU が Pentium4 2.8 GHz で，OS が FreeBSD 4.3-RELEASE の計

表 6　処理時間の評価結果
Table 6　Evaluation results of processing time.

	評価 1	評価 2	評価 3	文献 11)
処理 (1)	132.9 秒	534.7 秒	523.4 秒	20.8 秒
処理 (2)〜(5)	9.9 秒	339.0 秒	338.3 秒	3.8 秒

表 7　言語別の見逃しメール数
Table 7　Number of false negative emails of each language.

	3.4 節	4.2.4 項	評価 1（10 月）	評価 2（5 月）
日本語	61	31	114	362
非日本語	101	45	386	4,204

算機で行った．評価結果を**表 6** に示す．

処理 (1) は，受け取った迷惑メールから見逃しメールを抽出する処理である．評価 1 では 4,083 通の迷惑メールから 162 通の見逃しメールを，評価 2 では 13,916 通の迷惑メールから 4,566 通の見逃しメールを，評価 3 では 13,788 通の迷惑メールから 4,380 通の見逃しメールを，文献 11) の評価では 2,077 通の迷惑メールから 76 通の見逃しメールを抽出する処理である．処理 (1) は，抽出する処理を自動化するスクリプトを使用した．処理 (2)〜(5) は，処理 (1) で抽出した見逃しメールに対して，自動化スクリプトを実行したときの処理時間である．

表 6 の結果から，各処理にはメール数に比例した処理時間がかかることが分かる．また，見逃しメール数が 4,500 通程度の場合でも，処理 (2)〜(5) の処理時間は，約 340 秒であり，処理 (1) と合わせたとしても，約 15 分である．この処理を実行する頻度が短くても 1 カ月程度と想定すると，許容できる処理時間であると推察できる．ただし，処理 (1) は受信時に見逃しメールを分けて保存しておけば，省略可能である．

5.6　複数言語を扱う場合についての考察

提案方式は，見逃しメールについて言語を区別せずにトークンの累積度数分布を求め，判定に利用するトークンを限定する．このため，提案方式が特に有効に働くのは，見逃しメールに含まれるトークンが最も多い言語のメールに対してであると推察できる．各評価での言語別の見逃しメール数を**表 7** に示す．また，提案方式を適用して削減できた見逃しメールの言語別の内訳を**表 8** に示す．他の月と言語別のメールの割合は同様のため，評価 1 では 10 月，評価 2 では 5 月のメールを判定した場合を調査した．なお，評価 3 では提案手法で見逃しメールが減少しないため，この調査から除外した．

表 8　提案方式で削減した見逃しメール数
Table 8　Number of false negative emails reduced by proposed method.

	3.4 節	4.2.4 項	評価 1（10 月）	評価 2（5 月）
日本語	42	0	46	12
非日本語	33	9	162	287

3.4 節，4.2.4 項の評価では，日本語のメールが約 4 割程度である．表 8 から，3.4 節の評価では，日本語のメールについて特に効果が高いことが分かる．一方，4.2.4 項の評価では，非日本語のメールについてのみ見逃しメールを削減できたことが分かる．評価 1（10 月）については，日本語メールが約 2 割と少ないものの，それぞれの見逃しメールを 4 割程度削減できており，効果があることが分かる．評価 2（5 月）では，非日本語のメールが 9 割以上を占めているため，非日本語のメールでの削減率（6.8%）が，日本語のメールでの削減率（3.3%）より高い．これらのことから，言語別の見逃しメール削減数は，見逃しメールに多く含まれる言語について効果が高い傾向があることが確認できた．

6. 関 連 研 究

ベイジアンフィルタの実装方法として，POPFile[13] や Mozilla Thunderbird[14] のようにクライアント PC 上で動作するものと，bsfilter[10] や bogofilter[15]，SpamAssasin[16] 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．

ベイジアンフィルタの判定精度を向上させる方法として，様々な対策が提案されている．多言語環境における対策として，電子メールごとではなく，トークンごとに利用するコーパスを選択する方法[11] や言語や文字コードの知識を用いない方法[17] がある．文献 18) では，ユーザのフィードバックを用いて判定精度を向上させる手法が示されている．また，日本語処理において，トークン抽出法の比較や電子メールの迷惑メール確率計算に用いる単語数に上限を設けることの有効性が示されている[19]．文献 20) では，正当なメールに多く現れるトークンと迷惑メールに多く現れるトークンを同数選ぶことにより，判定精度を向上させる手法を提案している．さらに，文献 21) では，判定に利用するトークンの選択法について 4 つの方法を実験した結果を示している．これらの方法は，すべてのトークンを利用する方法，決まった数の特徴的なトークンを利用する方法，0.5 から設定した値よりも離れた迷惑メール確率を持つトークンを利用する方法，およびメールに含まれる全トークン数のうち設定した割合の特徴的なトークンを利用する方法である．いずれの方法も，提案方式とは異な

り，見逃しメールの原因に着目したものではない．

ヘッダ情報を有効に利用する対策として，迷惑メールに特徴的な傾向が現れる発信元情報を学習と判定に追加することで，ベイジアンフィルタの判定を補完する対策[22]や誤ったタイムゾーンやタイムスタンプ，および IP とドメインの不整合を利用した対策[23]がある．

また，利用者が自由にフィルタをカスタマイズできるように，フィルタ情報を可視化し，単語データや閾値を編集できるツールもある[24]．さらに，ベイジアンフィルタを単独で利用するのではなく，ホワイトリストやグレイリスト，またブラックリストと組み合わせた対策もある．リストの作成には，チャレンジ–レスポンスを用いた方法[25]や，メールの送受信関係の社会ネットワーク分析（SNA：Social Network Analysis）を用いた方法[26),27]がある．これらの対策は，広範囲の電子メールをカバーできるベイジアンフィルタと誤りが少ないリストを併用することで，両者の長所を生かし，短所を補う対策となっている．

以上のように，ベイジアンフィルタの精度を向上させるため，様々な対策がなされている．しかし，本論文で提案している見逃しメールに含まれるトークンの迷惑メール確率に着目し，判定に利用するトークンを限定することは行われていない．

7. おわりに

本論文では，ベイジアンフィルタにおいて，判定に利用するトークンを限定した迷惑メール対策の設計と評価について述べた．ベイジアンフィルタプログラムの 1 つである bsfilter について，判定したメールについて，その種類別に含まれるトークンの迷惑メール確率の分布を調査し，特に見逃しメールについてその原因を調査した．この結果，見逃しの原因の 1 つに，初めて出現するトークンの迷惑メール確率が低く，初めて出現するトークンが多くメールに含まれる場合があることを明らかにした．

そこで，この調査結果を基に，判定に利用するトークンを限定した迷惑メール対策を提案した．提案方式は，見逃しメールに含まれるトークンの迷惑メール確率の分布を調べ，見逃しの原因となりうるトークンが多く含まれる迷惑メール確率区間を迷惑メールの判定に利用しないことで，見逃しメールを減少させる方式である．また，提案方式により，判定に利用するトークンの迷惑メール確率の範囲を自動的に設定する手法について述べ，プログラムにより設定の自動化が可能であることを示した．

複数のメールセットにより評価した結果，提案方式により，誤検出メールを増やすことなく，見逃しメールを約 42%から約 6.5%減らすことができることを示した．ただし，初めて出現するトークンが見逃しの原因とならない場合は，デフォルトの設定と同じ方法で判定することになる．

今後の課題として，言語別に見逃しメールを調査し判定精度を向上させる方式の検討がある．

参考文献

1) The State of Spam: A Monthly Report - November 2010, available from ⟨http://www.symantec.com/content/en/us/enterprise/other_resources/b-state_of_spam_and_phishing_report_11-2010.en-us.pdf⟩.
2) フィッシング対策協議会：フィッシングレポート 2010, 入手先⟨http://www.antiphishing.jp/report/pdf/phishing_report_2010.pdf⟩.
3) Graham, P.: A Plan for Spam, available from ⟨http://paulgraham.com/spam.html⟩.
4) Robinson, G.: Spam Detection (2002), available from ⟨http://radio.weblogs.com/0101454/stories/2002/09/16/spamDetection.html⟩.
5) Robinson, G.: A Statistical Approach to the Spam Problem (2003), available from ⟨http://www.linuxjournal.com/article/6467⟩.
6) 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).
7) 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).
8) Wittel, G.L. and Wu, S.F.: On Attacking Statistical Spam Filters, *1st Conference on Email and Anti-Spam* (2004), available from ⟨http://www.ceas.cc/papers-2004/170.pdf⟩.
9) Lowd, D. and Meek, C.: Good Word Attacks on Statistical Spam Filters, *2nd Conference on Email and Anti-Spam* (2005), available from ⟨http://www.ceas.cc/papers-2005/125.pdf⟩.
10) bsfitler, available from ⟨http://bsfilter.org/⟩.
11) 岩永　学，田端利宏，櫻井幸一：ベイジアンフィルタリングを用いた迷惑メール対策における多言語環境でのコーパス分離手法の提案と評価，情報処理学会論文誌，Vol.46, No.8, pp.1959–1966 (2005).
12) TREC: TREC 2007 Public Corpus (2007), available from ⟨http://plg.uwaterloo.ca/~gvcormac/spam/⟩.
13) POPFile, available from ⟨http://popfile.sourceforge.net/⟩.
14) Mozilla Thunderbird, available from ⟨http://www.mozilla.com/en-US/thunderbird/⟩.

15) bogofilter, available from ⟨http://bogofilter.sourceforge.net/⟩.
16) SpamAssasin, available from ⟨http://spamassassin.apache.org/⟩.
17) 藤田拓也，松本章代，Martin J. Dürst：言語知識を用いないスパムメールフィルタに関する考察，情報処理学会研究報告，Vol.2008, No.122, pp.25–30 (2008).
18) Li, Y., Fang, B., Guo, L. and Wang, S.: Research of a Novel Anti-Spam Technique Based on Users? Feedback and Improved Naive Bayesian Approach, *Proc. International Conference on Networking and Services* (*ICNS'06*), pp.16–21 (2006).
19) 大福泰樹，松浦幹太：ベイジアンフィルタによる日本語を含むメールのフィルタリングについての考察，2006 年暗号と情報セキュリティ・シンポジウム（SCIS 2006）予稿集（CD-ROM）(2006).
20) 谷岡広樹，中川　尚，丸山　稔：特徴抽出方法の改善によるベイジアンフィルタの精度向上，情報処理学会論文誌：数理モデル化と応用（TOM），Vol.1, No.1, pp.175–184 (2008).
21) Deshpande, V.P., Erbacher, R.F. and Harris, C.: An Evaluation of Naive Bayesian Anti-Spam Filtering Techniques, *Proc. IEEE Information Assurance Workshop*, pp.333–340 (2007).
22) 伊藤朋哉，寺田真敏，土居範久：発信元情報を適用したベイジアンスパムフィルタ方式の提案，情報処理学会研究報告，Vol.2008, No.21, pp.285–290 (2008).
23) Chen, B., Dong, S. and Fang, W.: Introduction of Fingerprint Vector based Bayesian Method for Spam Filering, *5th Conference on Email and Anti-Spam*, available from ⟨http://www.ceas.cc/2008/papers/chenbin.pdf⟩.
24) 室伏　麗，齊藤泰一：ベイジアンフィルタの開発とフィルタ精度向上の研究，コンピュータセキュリティシンポジウム 2008（CSS 2008）論文集，pp.773–778 (2008).
25) 岩永　学，田端利宏，櫻井幸一：チャレンジ–レスポンスとベイジアンフィルタリングを併用した迷惑メール対策の提案，情報処理学会論文誌，Vol.45, No.8, pp.1939–1947 (2004).
26) 大福泰樹，松浦幹太：ベイジアンフィルタと社会ネットワーク手法を統合した迷惑メールフィルタリングとその最適統合法，情報処理学会論文誌，Vol.47, No.8, pp.2548–2555 (2006).
27) 白石善明，福田洋治，溝渕昭二，鈴木貴史：社会ネットワーク分析を用いたスパム対策：固有ベクトル中心性に基づくメールフィルタリング，情報処理学会論文誌，Vol.51, No.3, pp.1083–1093 (2010).

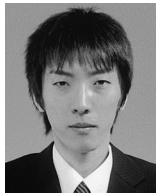
(平成 22 年 11 月 30 日受付)
(平成 23 年　6 月　3 日採録)

山内　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，USENIX 各会員．

上村　昌裕

2007 年岡山大学工学部情報工学科卒業．2009 年同大学大学院自然科学研究科博士前期課程修了．同年日本アイ・ビー・エム株式会社入社．コンピュータセキュリティに興味を持つ．



畑中　良太

2011 年岡山大学工学部情報工学科卒業．同年富士通エフサスシステムズ株式会社入社．コンピュータセキュリティに興味を持つ．