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

2007－ＣＳＥＣ－38 (46)
2007／7／20

特徴キーワードの変遷に見るSPAMメールの定性的特徴について

荒金 陽助† 佐野 和利† 塩野入 理† 金井 敦†

† NTT 情報流通プラットフォーム研究所 〒180-8585 東京都 武蔵野市 緑町 3-9-11

あらまし インターネットの普及に伴い，手軽なコミュニケーションツールとしての電子メールの利便性・重要性は無くてはならないものになってきている．しかしながら，電子メールの普及と呼応して，これを用いて強制的に情報を送りつけるspamメールが社会問題化してきている．spamメールに対しては，技術的および法的対策がなされており，特にspamメールフィルタリングはメールサーバの必須機能になってきている．本論文では1999年から2006年に届いた18931通の日本語spamメールを対象として，キーワードの変遷を調査した．その結果，変遷の特徴として減少型，復活型，凸型，増加型が存在すること，また，誘導的なspamメールが増加する傾向にあること，spamメールのキーワードは世の中のトレンドに従って別の同義語に変化してゆくこと，受信者のプライバシー意識を逆手に取ったキーワードが出現してきていることなどを明らかにした．

キーワード spamメール，電子メール，キーワード，傾向

Qualitative Characteristics of spam-mails Accoring to the Change of Their Keywords

Yosuke ARAGANE†, Kazutoshi SANO†, Osamu SHIONOIRI†,

and Atsushi KANAI†

† NTT Information Sharing Platform Laboratories, NTT Coporation
3-9-11 Midori-Cho, Musashino-Shi, Tokyo, 180-8585 Japan

Abstract According to spread of the internet, e-mail becomes useful and essential communication tool. However, respond to the popularization of e-mail, the spam mail which send non-solicited pornography and marketing e-mail is a social problem. Against the spam mail, there are many anti-spam technology and anti-spam act. Especially, spam filtering is an essential function in e-mail server. In this paper we analyze spam mail keywords of 18,931 spam mails received between 1999 and 2006. As a result, we could categorize the spam mails to decreasing type, revival type, hill type, and increase type. We find out some trends of leading spam mail is increasing, keywords change to other same meaning word according to public trend, and privacy sensitive keywords are increasing.

Key words spam mail, e-mail, keywords, change trends

1. はじめに

電子メールは，通信相手の時間を拘束することなく，瞬時に遠距離の通信相手に情報を送ることができ，そのコストも非常に低廉で，操作も簡単であるなどの多数の長所を有している．そのため，コミュニケーションツールとして電話を超える勢いで普及し，ビジネスシーンを始めとした重要な局面でも利用されるようになっている．また，我が国では携帯電話における電子メールの利用が非常に多く，また，非同期通信手段でありながらもリアルタイムに限りなく近い利用法が普及するなど，その利用形態もたいへん特徴的である．しかし，電子メールの普及に伴い，様々なセキュリティ脅威が顕在化してきている．コンピュータウィルスはコンピュータおよびネットワークに甚大な被害を及ぼすことが多く，セキュリティ脅威の代表的なものに挙げられている．コンピュータウィルスは様々な手段を用いて感染コンピュータの数を増やしてゆくが，最近のコンピュータウィルスのほとんどはネットワーク（特にインターネット）を介して感染を広げるようになってきている．このようなネットワークを介して感染を広げるコンピュータウィルスは，ターゲットとするコンピュータのソフトウェアの脆弱性を突いたり，電子メールによって感染ファイルを送りつけたりして別のコンピュータに感染するようになる．特に電子メールによる感染では，ウィルスが感染した添付ファイルをユーザが誤って実行してしまう場合も多く，爆発的な感染に結びつくことが少なく

ない．

一方，社会に与える影響として，spam メールの問題が注目されるようになってきた [1]．郵便が通信の主流の時代にも DM（Direct Mail）という広告手法が使われていたが，印刷および郵送のコストが発送量に比例して必要となること，発送する際には事業者の所在確認が行われることなどにより，一定の抑制がかかっていたと考えられる．故に，大規模の DM を発送する業者はコストの面からも非常に限定されていた．しかし電子メールを利用することで DM の形態が大きく変わることとなった．まずコストの面からは，従量制ではなく定額制であることから発送量に対するコスト的な歯止めがかからなくなったことなど，電子メールを使うことによるメリットが大きい．また，発送者の制限が無くなったことも spam メールの特徴である．インターネットは超分散環境であり，ネットワークの隅々まで全てのレイヤに渡る厳密な管理が行き届いているわけではない．そこで，匿名的にまたは他の国からネットワークに接続することで，発信者を特定することが困難な状況で spam メールを大量送信することが可能となっている．これによって不法な情報を多くのユーザに発送することも発生しており，ネットワーク帯域の圧迫と併せて，メールボックスが大量の spam メールで埋まってしまう問題や，グロテスクな内容など不快感を受ける spam(注1)の問題などが発生している．さらに，フィッシング詐欺などの犯罪に利用されることも特徴であるといえる [2]．

本稿では，この spam メールについて概観すると共に，その特徴として spam メールに含まれるキーワードに着目し，その変遷を抽出・考察する．以下，第二章では spam メールについて概観し，第三章にて spam メール調査手法を説明する．そして第四章で調査結果を考察し，第五章で本論文をまとめる．

2. spam メールについて

本章では，インターネット利用者に大きなインパクトを与えている spam メールについて概観する．

2.1 spam メールとは

spam メールが spam メールと呼ばれるようになったのはここ十数年に過ぎない．しかし，ジャンクメールやチェーンメールなどの名称で，以前より存在していることが知られている．1978 年にはインターネットの前身である arpanet にジャンクメールが送信されたことが知られている [3]．spam という名称は Hormel Foods 社の豚肉製品である SPAM [4], [5] に由来しているとされる．英 BCC のコメディ番組「Monty Python's Flying Circus」のレストランを舞台にした SPAM という作品において，SPAM を欲しない客に対しても強制的に SPAM を使った料理を売ろうとするやりとりが，受信者が欲しないメールを強制的に送るジャンクメールを指すようになった，という説が有力である．これは，米国における spam メール対策法である CAN-SPAM Act（Controlling the Assault of Non-Solicited Pornography and Marketing Act）の名称に如実に表れている．「要求されていないポルノや広告による攻撃」が spam メールの定義であるといえる．

2.2 spam メールの被害と対策

spam メールの発生以来その数は着実に増加を続け，現在ではインターネット上に流通する全電子メールのうちのほとんどを占めていると言われている [6]．その影響は大きく，溢れかえる spam メールの処理（spam メールの廃棄や spam メールに埋もれてしまったそれ以外のメールの抽出など）に必要となる膨大な時間，膨大な spam メールによるネットワーク帯域の圧迫など社会問題となっている．ヨーロッパの企業で 28 億ドル以上，アメリカの企業で 200 億ドル以上の被害があると言われている [3]．

このような spam メールに対して様々な対策がなされている．メール文面など特徴から spam メールを排除する spam メールフィルタや，spam メール発送者（spammer）のアドレスなどからのメールを排除するブラックリスト方式，送信ドメイン認証，Outbound Port25 Bloking など，様々な手法が考案され，利用されている [7]〜[9]．

一方，このような技術的解決策だけでなく，法律の側面からも spam メールの対策が行われている [10]．日本では，2002 年 4 月に可決成立し，同年 7 月 1 日より施行された迷惑メール規制 2 法（「特定電子メール送信適正化法」と「改正特定商取引法」）がある．これは，電子メールの表題に「未承諾広告※」と表示すること，連絡先を明記すること，拒否を意思表示した者に対して電子メールを送信することを禁止すること（オプトアウト）を義務づけている．欧米諸国においても同様の法律が施行されている．米国において 2004 年 1 月から施行された CAN-SPAM Act. では，オプトアウト制の採用と虚偽アドレスの表示禁止がなされている．欧州連合（EU）の「プライバシーと電子通信に関する指令」では，オプトイン制（送りつけに事前に受信者の同意を得ることを条件とする制度）が示されている．しかし，CAN-SPAM Act の施行直後と 2 年後の状況を調査した論文では，当初より法律が規定する表題の記述や，広告主の郵便住所，有効なオプトアウトなどの基準の遵守の程度は全体的に低かったが，2004 年に比べて 2006 年にはさらに低下していたという報告がなされており，spam メール問題の解決が容易でないことが示唆されている [11]．

3. spam メールの特徴調査

spam メールの特徴については，メールアドレスと受信する spam メールの量との関係や spam メールのターゲッティングセグメント，到着時間の特性など様々な調査がなされてきている [12]〜[14]．本論文では，日本語の spam メールについて，使われているキーワードがどのように変化してきたのか，その変遷を調査する．

3.1 対象とする spam メール

本論文では，10 名の研究者に協力を仰ぎ，1999 年から 2006 年までに届いた 18931 通の日本語の spam メールを提供してもらい，これを調査対象とした．この spam メールの各四半期（三ヶ月）ごとの特徴からその変遷を見ることにする．各四半

(注1)：精神的ブラウザクラッシャー（ブラクラ），マインドクラッシャー（マイクラ）などと呼ばれる．

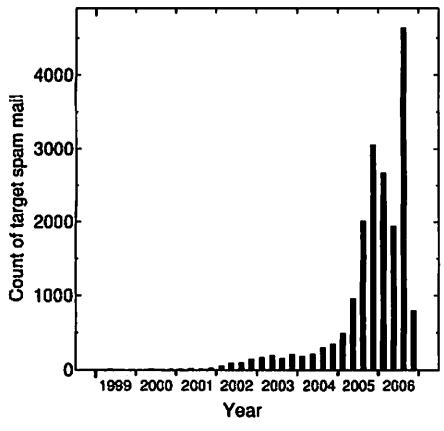

図 1　調査対象の spam メール数分布

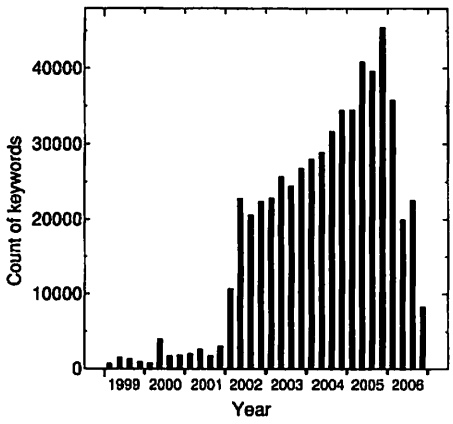

図 2　調査対象のキーワード数分布

期の調査対象 spam メールの数を図 1 に示す．

2001 年までは非常に少数であるが，2005 年以降は多数の spam メールが存在する．そこで後述する調査に当たっては，メール数による正規化を行うことでメール数の多寡の影響を排除するよう努めた．なお，受信者による偏りも存在するが，本稿での検討対象とはしない．

3.2 調査手法

本論文では，spam メールの特徴として，メール内の特徴的なキーワードの出現を用いた．プロセスは以下の通りである．

（1） キーワード抽出

全ての spam メールを形態素解析器 [15] にかけ，名詞および連続した名詞をキーワードとして抽出する．各四半期ごとに抽出されたキーワード数を図 2 に示す．

（2） TF/IDF 値の算出

各四半期の spam メールの本文をひとつの文書（メール）として，各キーワードのその四半期の出現回数 Tf_i を算出する．次に各キーワードをインターネットの検索エンジンで検索し，ヒットしたページ数を Df_i とする．これらを用いてそのキーワードが各四半期の spam メールをどれだけ代表するキーワードであるかどうかを示す TF/IDF 値 P_i を算出する．

$$P_i = Tf_i \times \log\left(\frac{N}{Df_i}\right) \tag{1}$$

ここで，N は使用した検索エンジンのインデックスが張られたページ数である 80 億とした．

（3） 特徴あるキーワードの抽出

各四半期において TF/IDF 値が上位 100 位となるキーワード 1164 語から，包含関係などの重複や記号が主体の無意味なキーワードなどを削除して，特徴あるキーワード 208 語を抽出する．

（4） 各四半期における順位の指標化

特徴あるキーワード 208 語について，図 2 に示す各四半期のキーワードにおける TF/IDF 値の順位を以下のように指標化する．なお，この指標化において図 2 に示す各四半期のキーワード数による正規化を併せて行う．キーワード i の四半期 m における指標値 V_{im} は，その四半期に抽出されたキーワード数 N_m と，その中での単語 i の TF/IDF 値の順位 O_{im} を用いて，以下のように定義する．

$$V_{im} = \log\left(\frac{N_m}{O_{im}}\right) \tag{2}$$

なお，キーワードがその四半期に一回も出現しない場合は，$V_{im} = 0$ とする．

（5） 考察

特徴あるキーワード 208 語それぞれについて四半期ごとの指標値の変遷を考察する．

4. 解析と考察

本章では，前章で説明した手法で解析を行った結果を示すと共に，それについて考察する．

4.1 解析結果

特徴あるキーワード 208 語について，1999 年最初の四半期から 2006 年最終の四半期までの 32 期間について指標値を算出した結果，平均 0.453，標準偏差 $\sigma = 1.26$ となった．この指標値の変遷において，大きく 5 つのパターンに分類された．

（1） 減少型

比較的単調に指標値が減少傾向にあるキーワード．「携帯電話」や「受験」など 8 つのキーワードが該当した．

（2） 復活型

過去に指標値が多いものの，一度減少した後，増加に転じたキーワード．復活の度合いが大きい（最大値の半分程度まで復活した）キーワードが「登録」や「配信停止」など 3 キーワード，大きな増加が見られないキーワードが「プレゼント」や「入力」など 4 キーワードであった．

（3） 凸型

増加していた指標値が減少に転じたキーワード．「返信」や「送信」など 6 キーワードが該当した．

（4） 増加型

比較的単調に指標値が増加しているキーワード．「好評」や「コチラ」などの漸増種 7 キーワードと「無料掲示板」や「仕事がら」などの急増種 33 キーワードが該当する．

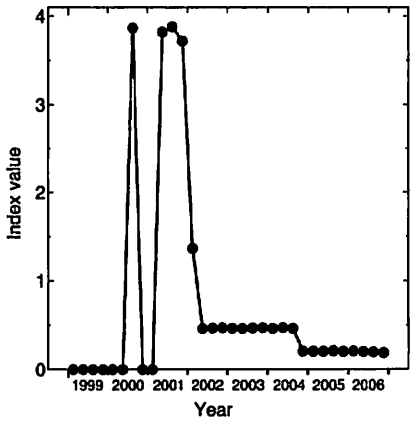

図 3　減少型のキーワード（携帯電話）

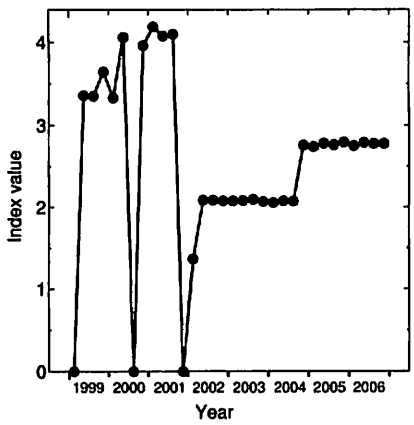

図 4　復活型のキーワード（登録）

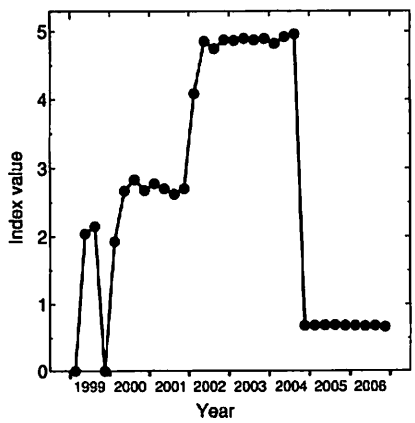

図 5　凸型のキーワード（返信）

（5）　その他

出現回数が極端に少なく，またばらけており，パターンに分類できないもの．本論文では考察の対象外とする．

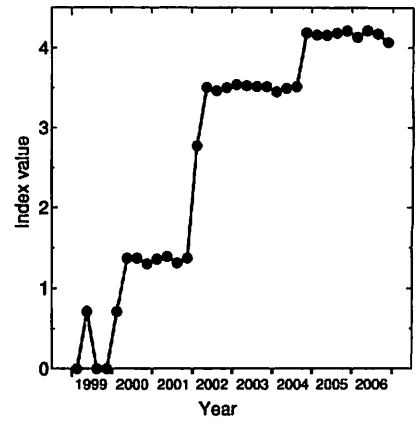

図 6　増加型のキーワード（好評）

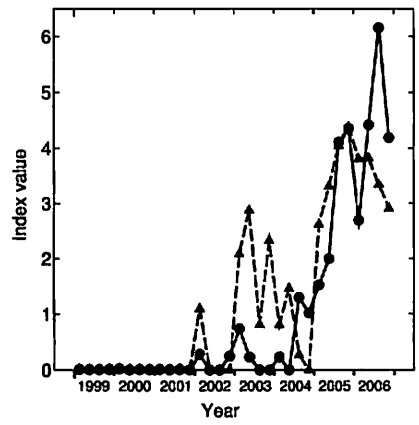

図 7　誘導キーワードの増加
配信不要：丸実線，配信拒否：三角破線

4.2 考　　察

各キーワードの指標値の推移を比較して見えてきた事象を以下に示す．

- 取引から誘導へ

双方向的なインタラクションを期待する「取引型」のキーワードは，登録（復活型），送付希望（減少型），返信（凸型），注文（増加型漸増種），価格（増加型漸増種）など全期間を通して高い指標値を示すキーワードが存在する．広告ないしはそれと思わせる対価型の spam メールが，形を変えつつも継続的に存在していると思われる．

一方，配信拒否や配信不要，退会などの spam メールの受信を拒否するための操作から，有効なメールアドレスの情報などを取得するための誘導と思われるキーワードは増加型急増種がほとんどである．さらに，Web サイトなどへの誘導を行う，サイト登録，登録画面，利用案内などのキーワードも増加型急増種である．また，これらのキーワードと組み合わせる「コチラ」というキーワードも同様に増加型急増種である．

- 言葉の変遷

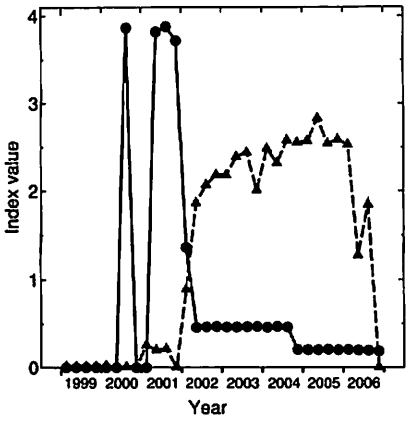

図 8　キーワードの変化（携帯電話→モバイル）
携帯電話：丸実線，モバイル：三角破線

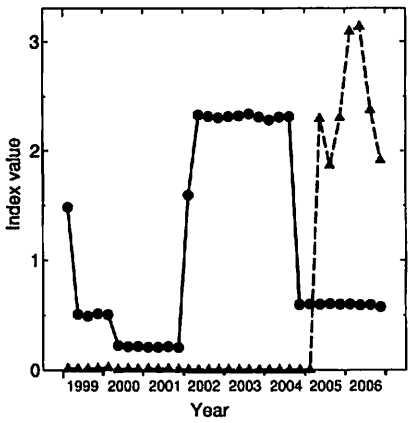

図 9　キーワードの変化（アドレス→メアド）
アドレス：丸実線，メアド：三角破線

同様の意味を持つ言葉であっても，時間と共に使われる割合が変化している．携帯電話や携帯が減少型であるのに対し，モバイルは増加型漸増種である．また，アドレスが凸型であるのに対し，メアドは増加型急増種である．spammer が受信者の心を捉えるために，トレンドに合致したキーワードを利用している結果である．これらの要因としては，上記のような能動的な理由もさることながら，ベイジアンフィルタリングなどによる spam フィルタに spam メールのキーワードとして登録されることにより，以後は使われなくなったことも考えられる．また，無料掲示板などのキーワードが増加型急増種として 2005 年半ばから現れており，今後も世の中の情勢を追いかけた新しいキーワードが spam メールに出現してゆくことが予想される．

- プライバシー意識

受信者のプライバシー意識に訴えかけるキーワードが増加してきていることも特徴であると考えられる．2004 年頃は個人認証といったサービス加入に際するセキュリティキーワードが出現していたのに対し，2005 年半ばからは完全匿名，一切関与，安心などのキーワードの指標値が急増してきている．

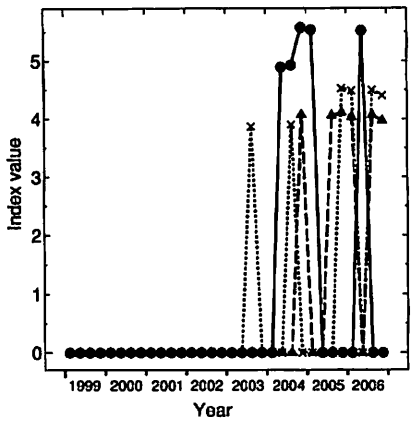

図 10　セキュリティに関わるキーワードの変化
個人認証：丸実線，一切関与：三角破線，完全匿名：×点線

なお，一切関与というキーワードは，メール本文内にて大きく二通りの使い方がなされている．ひとつは「当社は架空請求・ワンクリック詐欺等には一切関与しておりません」という安全性をアピールする使い方であり，もうひとつは，「女性会員様との間での会話等，プライバシーについては当クラブは一切関与しておりませんので，ご安心して存分にお話下さいませ」というプライバシー尊重によって利用を促進させる使い方である．

このようなサービス利用者間の直接のやりとりを想起させるようなキーワードは他にも使われており，その多くが増加型急増種である．これら，女性会員，アドレス交換などのキーワードによって，サービスのメリットを強調すると共に，女性誌や会員数などのキーワードによってサービスを享受できる可能性が高いことを謳っていることも特徴であるといえよう．

5. おわりに

本論文では，spam メールおよびそれへの対策について概観すると共に，特徴的なキーワードに着目して 1999 年から 2006 年までの 18931 通の spam メールの特徴調査を実施した．各キーワードの TF/IDF を求めることで，spam メールを特徴付けると考えられる 208 語を抽出し，この 208 語について四半期ごとの出現指標値の変化を観察した．その結果，キーワードの変遷は減少型，復活型，凸型，増加型に大きく分類できることが分かった．これらのキーワードのいくつかを抽出して考察し，誘導的な spam メールが増加する傾向にあること，spam メールのキーワードは世の中のトレンドに従って別の同義語に変化してゆくこと，受信者のプライバシー意識を逆手に取ったキーワードが出現してきていることなどを明らかにした．

本論文では，10 名の受信者に届いた spam メールを対象としたが，受信者は全てセキュリティ研究者であり，母数とした spam メールに偏りが生じていることも考えられる．今後は，さらに一般的な母数の spam メールを対象とした調査を行う必要があるだろう．また，解析・考察に当たっては，TF/IDF を

用いた指標値を導入したが，キーワードの遷移の分類においては定性的な評価にとどまった．今後の課題として定量的な評価手法を検討する必要がある．

文　献

[1] 山井成良, 桝田秀夫, "spam メールの現状と対策の動向", 情報処理, Vol.46, No.7, pp.739–740, 2005.

[2] 荒金陽助, 間形文彦, 柴田貫介, 塩野入理, 金井敦, "フィッシング詐欺対策に関わる研究動向と法適用について", 情報処理学会マルチメディア, 分散, 協調とモバイル (DICOMO2006) シンポジウム, 6H2, pp.777-780, Jul. 2006.

[3] Stephen Hinde, "Spam: the evolution of a nuisance", Computer Security, Vol.22, No.6, pp.474–478, 2003.

[4] Hormel Foods Corporation Web Site, http://www.hormel.com/

[5] SPAM Web Site, http://www.spam.com/

[6] 総務省 迷惑メールへの対応の在り方に関する研究会，迷惑メールへの対応の在り方に関する研究会 最終報告書，http://www.soumu.go.jp/s-news/2005/pdf/050722_2_02_00.pdf, 2005.

[7] Paul Graham, "A Plan for Spam", http://paulgraham.com/spam.html, Aug. 2002.

[8] 王暁, 堀良彰, 桜井幸一, "中国語迷惑メールにおけるベイジアンフィルタの適用と評価", 情処研報, 2006-CSEC-33(9), pp.45–50, May. 2006.

[9] 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", 情処論, Vol.47, No.8, pp.2548–2555, Aug. 2006.

[10] 岡村久道, "サイバースペース法律相談所　第 3 回　迷惑メールの法的規制", 情報通信ジャーナル, Vol.22, No.7, pp.24–25, Jul. 2004.

[11] Galen A. Grimes, "Compliance With the CAN-SPAM Act of 2003", Communications of the ACM, Vol.50, No.2, pp.56-62, Feb, 2007.

[12] TanZila Ahmed and Charles Oppenheim, "Experiments to identify the causes of spam", Aslib Proceedings New Information Perspectives, Vol.58, No.3, pp.156–178, Mar. 2006.

[13] Il-Horn Hann, Kai-Lung Hui, Yee-Lin Lai, S.Y.T. Lee, I.P.L.Png, "Who Gets Spammed?", Communications of the ACM, Vol.49, No.10, pp.83–87, Oct. 2006.

[14] 市川貴久, 奥田隆史, 井手口哲夫, "ケーススタディによる spam メールの到着間隔特性の解析", 信学技報, Vol.206, No.524, IN2006-172, pp.59-64, Jan. 2007.

[15] 形態素解析エンジン MeCab, http://mecab.sourceforge.net/.