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

2007－ＣＳＥＣ－38 (48)
2007／7／20

迷惑メールの外見的特長についての一考察

神谷 造† 柴田 賢介† 佐野 和利† 荒金 陽助† 塩野入 理† 金井 敦†

†日本電信電話株式会社 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
E-mail:

†{kamiya.itaru, shibata.kensuke, sano.kazutoshi, aragane.yosuke, shionoiri.osamu, kanai.atsushi}@lab.ntt.co.jp

あらまし インターネットの普及に伴い，電子メールという最も手軽に利用できるコミュニケーション手段において，迷惑メールの存在が問題になってきている．迷惑メールの多くは任意のサイトに誘導するのを目的としており，効果的にそれを実現するためにメールの本文が作成されている．その際しばしばメール本文の文章の再利用を行うことがある．本研究では迷惑メール送信者がどのように迷惑メールを作成するのか，迷惑メール 15 万通を対象にメールと誘導先サイトの関係とメール本文の同一性の調査を実施した．本論文では，調査内容および得られた知見と考察について述べる．

キーワード 迷惑メール，電子メール，再利用

A study of outward features of Spam mails

Itaru KAMIYA†, Kensuke SHIBATA†, Kazutoshi SANO†, Yosuke ARAGANE†, Osamu SHIONOIRI†, and Atsushi KANAI†

†NTT Information Sharing Platform Laboratories, 3-9-11 Midoricho Musashino-Shi Tokyo 180-8585 Japan
E-mail:

†{kamiya.itaru, shibata.kensuke, sano.kazutoshi, aragane.yosuke, shionoiri.osamu, kanai.atsushi}@lab.ntt.co.jp

Abstract The purpose of spam mail sending is mostly leading spam mail readers to some web sites. In order to lead efficiently, spam mail senders create attractive writings for mail body. On this occasion, spam mail senders often use legacy writings which they used as spam mails' body before. In this paper, we show the results of analysis on hundred fifty thousand spam mails about leading sites and frequency of reusing former spam mails' writings.

Keyword spam mail, e-mail, reuse

1. はじめに

日本におけるインターネットの普及は目覚しく，2005 年の時点では総人口の 66.8%に相当する 8529 万人がインターネットを利用し，企業においては 97.6%が利用しており，社会のインフラとして必要不可欠なものとなっている[1]．その中で電子メールは最も古くから利用され，その非同時性，記録性，同報性等と，手軽に利用できる環境が実現されていることから広く利用され，重要な通信手段となっている．

しかし，電子メールの利便性は利用者の活動を効率化する一方で，迷惑メールの蔓延という影の部分も生み出している．これは電子メールの宛先さえ正しいものを設定しれば，その宛先に対してメールを送信できるという電子メールの特徴により実現されるものである．これと同様のことは一般の郵便でも実現されているが，その送信のための費用が郵便に対して比較にならないほど小さく実効的に 0（ゼロ）であることが電子メールにおける迷惑メールの原因ともなっている．2006年後半では電子メールトラフィックの6割が迷惑メールである[2]という報告も存在する．

電子メールのもう一つの大きな特徴として，送信者の特定の困難さがある．ヘッダには送信者情報や配送情報を持つことが可能であるが，その情報は容易に偽装を行うことができ，送り手は自分自身への追跡から容易に身を隠して電子メールを送ることができる．これにより，悪意を持った送信者は，単に受信者に迷惑となる電子メールを送りつけるだけではなく，電子メールを利用してマルウエアの媒介を行ったり，フィッシング詐欺のための誘導を行ったりする等のサイバー犯罪を実行することも可能となっている．

このような電子メールの悪意の利用の第一歩となる迷惑メールへの対策は，日本では「特定電子メールの送信の適正化等に関する法律」[3]による法的な規制を行っており，海外でも，米「CAN-SPAM 法」，EU「電

子通信分野における個人データ処理及びプライバシー保護に関する指令」(Directive 2002/58/EC)，英「2003年プライバシー及び電子通信（ＥＵ指令）規則」，韓国「情報通信網利用促進及び情報保護等に関する法律」等の法規制を実施しているが，先に述べたように送信者は自己への追跡から容易に身を隠すことがきるため，これら法規制も実効的な効果をあげることができていない．

このため，迷惑メールに関しては利用者が様々なセキュリティソフトウエアやISPの助けを借りて，主としてフィルタリング技術[5]-[9]による迷惑メールの振り分けを行い，その被害から逃れようとしている．しかし，フィルタリングではメール送信者側もフィルタリングに引っ掛からないようにメールの内容を変更し，あるいは，画像を用いてフィルタリングをすり抜けるような迷惑メールの高度化をはかり，その対策はいたちごっことなっているのが現状で，今後とも継続的なフィルタリング技術の開発は必要とされている．

本研究は迷惑メールのフィルタ技術に寄与する新たな観測点の追加を目的として，迷惑メールの誘導先サイトとの関係と電子メールの本文の同一性に着目し，迷惑メールの調査を行った．

2章では調査内容について説明し，3章にて調査結果を示す．4章ではそれら調査結果から考察を行い，5章をまとめとする．

2. 調査

今回の調査では，迷惑メールは受信者に本文を読ませ，サイトへ誘導することを目的としているものが多いため，メールの本文に記載される誘導先であるURLリンクと本文そのものを調査対象とした．調査対象としたメールは研究開発者のアドレス宛に送られたもので，各アドレスの所有者により迷惑と判定されたメールを使用した．迷惑メールはアドレス数18個，対象となる受信期間は1996年から2006年で，総数は150,233通であった．

2.1. リンク情報の抽出

分析は各迷惑メールを機械的にスキャンして以下の内容を抽出した．

・メール受信日時
　メールヘッダより抽出した．メールヘッダには送信日を表すdateフィールドがあるが，送信者側で詐称可能であるため，今回Receivedフィールドに記載されている直近のSMTPサーバの受信日時をメール受信日時として抽出した．

・記載リンク
　メール本文中に現れるURLを抽出した．URLには受信者の特定にも使用されるビーコンがパラメータとして付いていることがあるため，パラメータがURLについている場合は，それらを切り離したものをリンクとした．なおリンクの内HTMLメールの画面オブジェクト(gif, jpg)と思われるものは集計から除外している．

2.2. 類似メールの抽出

調査対象のすべてのメールの本文の比較を行い本文文章が一致するメールを類似メールとして抽出した．一致判定するに当たっては以下の手順で実施した．

1. メールからボディ部分を抽出
2. 1で抽出したものから改行のみの行，記載メールアドレス，記載URLを削除し比較対象の本文を作成
3. HTMLメールについては表示テキストを抽出した後に2の処理を実施
4. 2，3で加工した本文を全て比較し完全一致するものを類似メールとして抽出

3. 調査結果

分析メール数150,233通に対し調査を実施した．調査結果を以下に示す．

3.1. 記載リンク数

メール総数150,233通に対し，リンクの記載されていないメール数は34,044通あり，これを除外したメール116,189通に対して集計した．記載されていた一意なリンク数は150,268リンク観測された．またリンクそれぞれの迷惑メールへの累計記載数は357,998回であった．

リンクの出現状況について，1メール毎のリンク数の集計結果を図1に示す．

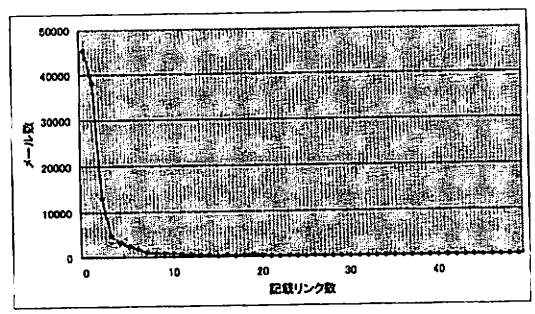

図1:メール毎のリンク数

メール毎の平均リンク数は3.08リンク/mailであった．迷惑メール全体で記載リンク数がどの様な割合になっているのかを見るために，図2にリンク数毎の累計メール数の割合を集計した．

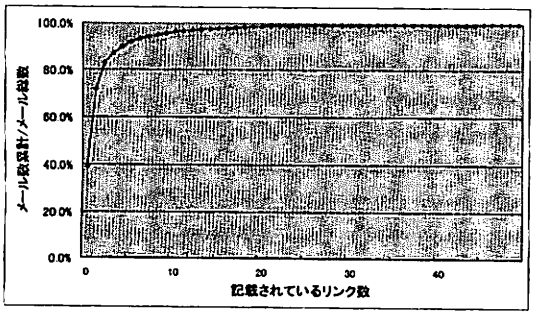

図 2:リンク数毎の累計メール数の割合

メール全体の 90.0%のメールでは記載されるリンク数は 5 リンク以下であった.

3.2. 同リンクのメール記載状況

リンクの記載期間を見るため，そのリンクが初めて出現した迷惑メールの受信時刻と最後に出現した迷惑メールの受信時刻の間隔をリンクが記載されていた期間とし，記載期間毎にリンク数を集計した．集計結果を図 3 に示す.

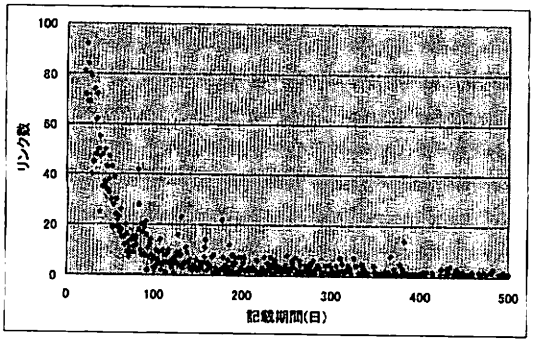

図 3:リンクの記載期間

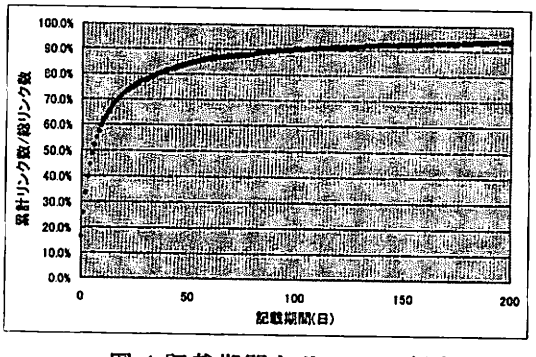

図 4:記載期間とリンクの割合

リンクの平均記載時間は 55.73 日で，記載期間が長くなるにつれリンク数がなだらかに減少していく傾向があった.

記載期間に対する累計リンク数の全体に対する割合を図 4 に示す．リンクの 61.7%が記載期間 10 日間以内であり，記載期間 100 日以内のリンクは全体の 90.2%であった.

リンクの 61.7%が記載期間 10 日間以内であり，記載期間 100 日以内のリンクは全体の 90.2%であった.

3.3. 類似メール数

メール総数 150,233 通の本文として観測された数は 107,945 種類で，そのうち 14,972 種類は 2 通以上のメールで観測された．他のメールと本文一致したメールが観測された数を類似メール数とし，類似メール数ごとのメール数の集計を図 5 に示す.

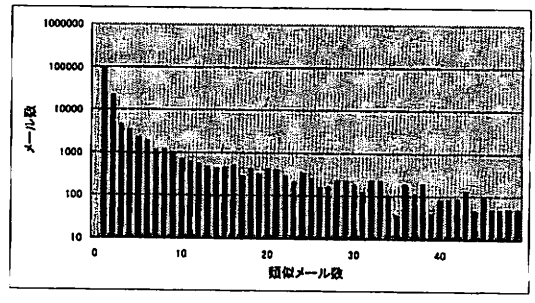

図 5:類似メール数毎のメール数

全く他のメールと類似しなかったメール(類似メール数:1)は 92,973 通存在した．一方類似していると判定されたメール数は 57,260 通観測され，全メールの 38.1%をしめた.

3.4. 類似メールの月別集計

類似と判定されるメールの割合の時間変化を見るために，他に同じ文面のメールが存在するメール(類似メール)数と他に同じ文面のメールが存在しないメール(単独メール)数の月別集計を図 6 に示す．また月別の類似メールの割合を図 7 に示す.

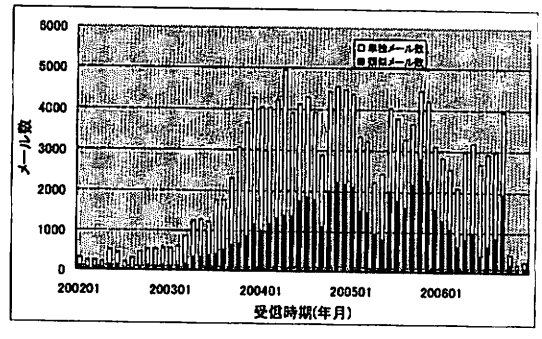

図 6:月別の類似メール観測数

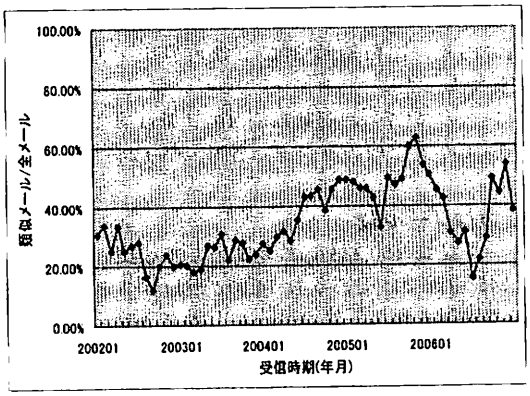

図 7:月別の類似メール数の割合

時期により割合が 10%強から 60%強までのばらつきがあった.

3.5. 類似メールの平均受信間隔

類似メールがどのように使用されているかを知るために類似メール毎に平均の受信間隔を集計した. 日単位の平均受信間隔に対する類似メールの種類数(類似グループ数)を図 8 に示す.

また平均受信間隔に対する累計の類似グループ数の全類似グループ数に対する割合を図 9 に示す.

類似グループ数総数 14,973 種類のうちすべてが同時刻に発信されているものは 5,247 種類 35.0 %を占めていた. また平均受信間隔が 1 日以内のものは 10,869 種で全体の 73%を占めていた.

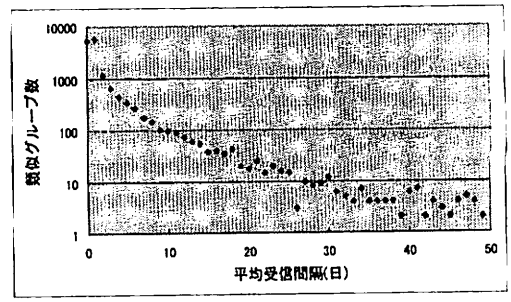

図 8:平均受信間隔と類似グループ数

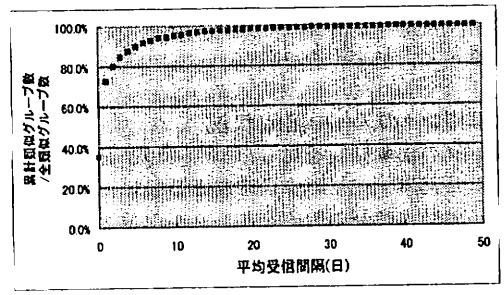

図 9:類似グループ数の割合

3.6. 類似メールの受信時間の分布

同時発信以外の類似メールについてその発出時刻のばらつきの程度を見るため類似グループ毎に標準偏差を集計した.

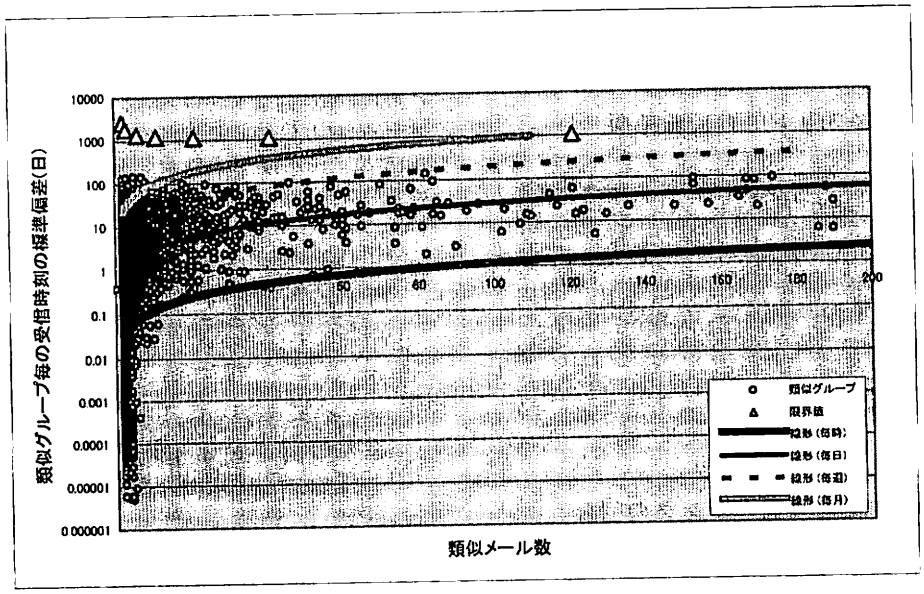

図 10:類似グループ毎の送信時刻の標準偏差と類似メール数

各類似グループにおける類似メール数と標準偏差の集計結果を図10に示す.各類似グループのプロットは類似メール数50通の領域に集中している.50通以上の傾向を受信間隔と比較するためにコンスタントに等間隔で受信した場合の期待値を図中に線で示した.50通以上では各類似グループのプロットは1時間から1週間の範囲に収まる傾向があった.なお,メール集計期間が10年と限られた期間であったため,標準偏差の値には上限値が存在する.毎月10年間類似メールを受信すると類似メール数は120通になる.これについても図中に限界値としてあわせて表示した.

4. 考察

4.1. 観測されたリンクについて

リンクについては同じリンクがメールに記載されて届く期間はあまり長くなかった.フィッシング詐欺を行うサイトにかぎって言えば寿命は4日以内[9]であるが,今回の調査ではメールの目的を特に固定せず広く迷惑メールを調査対象としたにもかかわらず10日以上記述されないリンクが全体の60%も存在した.迷惑メール送信者はサイト運用者から広告・宣伝としてメールを発送する業務を請け負っていると考えられるが,その際請負う期間が数日のオーダーである可能性が考えられる.また90%のリンクが100日以上記載されないという点では,繰り返し迷惑メールに記載されても誘導対象となるリンク先のサイトの寿命が数ヶ月のオーダーである可能性が考えられる.

4.2. 本文の同一性について

一方本文の再利用性については,受信者に表示される文面の完全一致という条件であったにもかかわらず,38.1%の迷惑メールが一致していた.本文の文面が一致したメールのうち35.0%は同時に発出されていたと考えられるが,残りは発出時刻にずれがあり,送信者側にて文面の再利用が行われたと考えられる.これら時間的ずれが確認できた同一メールの標準偏差をみるとほぼ1時間から1週間以内におさまっている.故に受信者が初受信時に迷惑メールとして振り分ければあとは単純なパターンマッチでフィルタすることが可能であり,今回の迷惑メールに関して言えば24.8%(0.381×(1-0.350))存在するこのような迷惑メールを振り分けることが可能と推測される.ただし,メール本文が完全一致したメール数は図7のとおり時期に応じて大きく変化するため,過去に迷惑メールと判定されたメールの本文と完全一致するメールを除くというフィルタリング手法を適用しても常時一定の効果を望むのは困難である.

調査後,類似メール判定されたメール本文を他の類似メールの本文と比較してみたところ,パラグラフ単位で文章が一致していたり,1文・1単語のみ異なっていたりするケースが確認できた.部分的な文章一致を考慮すればフィルタリングの効果を今後大きくしていくことも可能であると考えられる.

5. まとめと今後

迷惑メールのフィルタ技術に寄与する新たな観測点の追加を目的として,迷惑メールの誘導先サイトと電子メールの本文の同一性について,迷惑メールの調査を行った.調査ではリンクの迷惑メールへの記載期間が短いこと,本文再利用が行われるメール数が25%程度存在することを明らかにした.今後はメール本文の部分一致状況の分析を進め,この観点での迷惑メール対策の検討を進めていく予定である.

文　献

[1] 総務省,情報通信白書 平成18年版, http://www.johotsusintokei.soumu.go.jp/whitepaper/whitepaper01.html, 2007.

[2] Symantec社,シマンテックインターネットセキュリティ脅威レポート Volume XI: 2007年3月, http://www.symantec.com/content/ja/jp/enterprise/white_papers/wp_istr11_2007.pdf, 2007.

[3] 総務省,特定電子メールの送信の適正化等に関する法律, http://www.soumu.go.jp/joho_tsusin/top/pdf/meiwakulaw_h17.pdf, 2002.

[4] 原田透,中野学,松本勉,"メール不達通知を利用した迷惑メール対策",情報処理学会シンポジウム論文集, Vol.2004, No.11, Page.1A-2, Oct. 2004.

[5] 長谷川明生,"Spamメールの解析",情報処理学会研究報告, Vol.2004, No.77, pp.37-42, Jul. 2004.

[6] 市川貴久,奥田隆史,井手口哲夫, Xuejun Tian,"ケーススタディによるspamメールの到着間隔特性の解析",電子情報通信学会研究報告, Vol.106, No.524, pp.59-64, Jan. 2007.

[7] 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, pp2548.2555, Aug. 2006.

[8] 野原史朗,志田晃一郎,横山孝典,"送信者情報を利用した迷惑メール対策",情報処理全国大会講演論文集, Vol.69th, No.3, pp3.351-3.352, Mar.2007

[9] Anti-Phishing Working Group, "APWG Phishing Trends Activity Report for April, 2007", http://www.antiphishing.org/reports/apwg_report_april_2007.pdf, 2007