統計的フィルタリングに対する Word Salad 攻撃についての考察

岩永 学 †　　田端 利宏 ‡　　櫻井 幸一 ‡

† 九州大学　大学院システム情報科学府
812-8581 福岡市東区箱崎 6 丁目 10-1
iwanaga@itslab.csce.kyushu-u.ac.jp

‡ 九州大学　大学院システム情報科学研究院
812-8581 福岡市東区箱崎 6 丁目 10-1
{tabata,sakurai}@csce.kyushu-u.ac.jp

あらまし　迷惑メール対策手法のひとつに統計的フィルタリングがあるが，この手法を回避するために最近，迷惑メール内に内容とは無関係な単語を挿入する Word Salad とよばれる攻撃が行われるようになっている．本発表では日本語の迷惑メールにおける Word Salad 攻撃の影響について実験を行った．実験の結果，日本語の迷惑メールにおいても Word Salad 攻撃が統計的フィルタリングによる迷惑メールの検出に影響を与えうることがわかった．

An analysis of word salad attack against statistical filtering

Manabu Iwanaga†　　Toshihiro Tabata‡　　Kouichi Sakurai‡

†Graduate School of Information Science and Electrical Engineering
‡Faculty of Information Science and Electrical Engineering
Kyushu University
6-10-1 Hakozaki, Higashi-ku
Fukuoka Fukuoka, 812-8581 Japan
iwanaga@itslab.csce.kyushu-u.ac.jp, {tabata,sakurai}@csce.kyushu-u.ac.jp

Abstract　Statistical filtering is one of major anti-spam methods. Recently, spammers have began inserting random words to their spam, called word salad attack, to evade filtering. In this paper we made trial of effect of word salad attack in Japanese email environment. The trial showed that word salad attack can be effective against statistical filtering in Japanese email environment.

1　はじめに

近年の電子メールの普及に伴い，送信に要する費用の少なさから迷惑メールが増加している．各国で迷惑メールの法的な規制のための取り組みも行われているが，十分な効果を挙げているとは言いがたい．

迷惑メールの問題点として，受信者の時間や費用を浪費することが挙げられる．大量に送られてくる迷惑メールは，受信者をいらだたせ時間や通信料金を浪費するだけでなく，迷惑メール以外のメールを見落としたり，誤って削除したりする危険性さえ生じさせる．また，通信事業者の側から見た場合，迷惑メールによるネットワークトラフィックの増大による，設備投資コストの増加という問題がある．

迷惑メール送信者による送信の手口は巧妙化しており，文面の巧妙化のほか，送信者アドレスの詐称をはじめとした電子メールのヘッダ等の偽造も行われている．このため，ヘッダの内容に対する単純な検査だけでは，迷惑メールを十分に検出することは難しい．

2　既存の迷惑メール対策手法

迷惑メールの増加に伴い，MUA（電子メールクライアント）やISP（インターネット接続業者）による迷惑メール対策も活発になりつつある．いくつかのMUAは迷惑メールに対するフィルタリング機能を標準機能として備えるようになった．また，ISPが到着する電子メールについて迷惑メールのフィルタリングを行ったり，独自のMUAにISPへの迷惑メール報告機能を設けてフィルタリングに反映することも行われている．

メールの文面によるフィルタリングのうち最近盛んに用いられているのが，統計的な手法を用いたフィルタリングである．これについては3章で詳しく述べる．

DNSを用いてメールアドレスを詐称した電子メール送信を防止し，迷惑メールの送信を防止する手法も提案されている [1]．これは，迷惑メールは送信者のメールアドレスを偽装することが多いため，メールアドレスの詐称を防止することで迷惑メール送信を困難にするという発想に基づく手法である．

電子メールをやりとりする通信相手ごとに異なるメールアドレスを生成するという方法もある [2]．通信相手の重要度によってメールアドレスを使い分けるという手法は従来から行われてきたが，この手法を機械的に補助し，自動的に異なるメールアドレスを生成するものである．通信相手1人ごとに異なるメールアドレスを作成すれば，そのメールアドレスが迷惑メール送信者に知られ，そのメールアドレスを破棄する場合にも新しいアドレスを知らせる必要のある相手は1人だけで済む．

送信者に手動での確認作業を求めたり，手間のかかる計算を行わせることによって大量送信を防止するという手法も研究され，また一部は実際に用いられている [3, 4, 5]．Computational PuzzleやReverse Turing Testと呼ばれる手法がこれに含まれ，迷惑メール送信者が単位時間当たりに送信可能な電子メールの数を大幅に減少させることで，1通あたりのコストを上昇させ，迷惑メール送信を採算の合わないものにするという意図に基づく手法である．

3　統計的フィルタリング

統計的なフィルタリングは導入の簡単さと比較的よい精度から近年人気を集めている．統計的フィルタリングは，Naive Bayes, Support Vector Machine (SVM)，Boosting，Markov Chainなどの手法を用いて，過去に存在した正当な電子メールや迷惑メールから抽出した特徴と比較し，判定対象の電子メールが迷惑メールか否かを判定するものである．このうち，Naive Bayesを用いる手法はベイジアンフィルタリングと呼ばれ，近年Grahamによる “A plan for spam[6]” をきっかけに数多くの実装が開発・公開されるようになった．これらの実装の中には，電子メールのメールサーバ到着時にprocmail[7]等を経由して呼び出され，サーバ上で動作するものや，Mail User Agent (MUA)がメールサーバにアクセスする際にクライアント上でプロキシサーバとして動作するもの，また既存のMUAの機能として組み込まれた形で動作するものなどが存在する．

3.1　統計的フィルタリングの回避

統計的フィルタリングが多くの利用者やソフトウェアに採用されるにつれて，迷惑メール送信者も統計的フィルタリングの回避を重視するようになってきた．英文の迷惑メールにおいては，多くが既に何らかのフィルタリング回避策をとるようになっている [8]．主な回避策には以下のようなものがある．Graham-Cummingは [8] で主な回避策について分析を行った．

- ランダムな単語や文章の付加
- 文章を画像形式で記述
- HTMLメールにおいて，プレーンテキストパートとHTMLパートに異なる内容を記述する
- 単語間の空白の代わりに記号等を用いる．また，単語内の文字を記号で分断する

本論文ではこのうち，最近の多くの迷惑メールにおいて行われている，ランダムな単語や文章の付加 (Word salad) に注目する．

Word salad は，迷惑メールの目立たない場所にランダムな単語や文章を付加し，それらに含まれる単語の一部が受信者の学習データ内にある迷惑メール確率の低い単語に一致する可能性を狙った攻撃である．迷惑メール確率の低い単語は電子メールに対する迷惑メール確率を低下させ，受信者の迷惑メールフィルタに正当な電子メールと誤認させる可能性がある．

多くの場合，Word salad は本文の末尾にランダムなアルファベット列や，辞書からランダムに選んだ単語を 10〜100 個付加する方法をとっている．[9] によれば，単語を付加する Word salad には，単純にランダムな単語を付加する Dictionary Word Attack のほかに，付加する単語から迷惑メールの特徴とみなされる可能性のある単語を除くことでより効果的に統計的フィルタリングの回避を図る Common Word Attack がある．[9] では小さな迷惑メール本体にランダムな単語を多数付加することによって，ナイーブベイズを用いたフィルタリングでは大半の迷惑メールを見逃してしまう場合があることが指摘されている．

英語においては単語の区切りとして空白が存在するため，英語の電子メールに対する統計的フィルタリングでは，通常，空白や一部の記号で区切られた単語をそのままトークンとする．それに対して日本語では単語間に空白が存在しないため，日本語の電子メールに対する統計的フィルタリングでは，トークンの抽出方法は実装により様々である．

たとえば，Mozilla[10] は連続するひらがな，連続するカタカナ，漢字 1 文字をトークンとして抽出する．Bsfilter[11] は連続するカタカナ，単独で存在する漢字 1 文字，隣り合う漢字 2 文字をトークンとして抽出する．Scbayes[12] では隣り合う日本語 2 文字のうち，1 文字目がひらがな・カタカナで 2 文字目が漢字となる組み合わせ以外のものをトークンとして抽出する．また，POPFile[13] などでは外部ソフトウェア（KAKASI[14]）によって単語ごとに分割し，分割した単語をトークンとして抽出する．これらの手法を図 1 に示す．

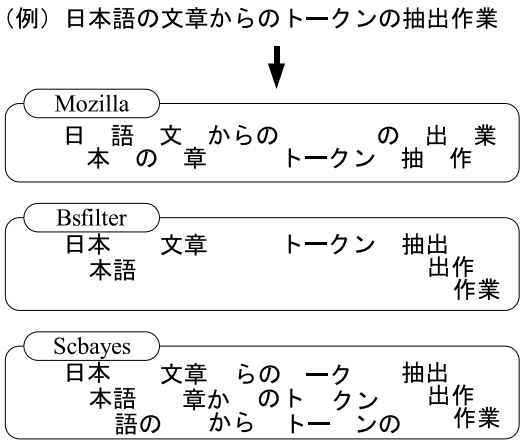

図 1: トークンの抽出方法

日本語の迷惑メールにおける Word salad 攻撃もその効果や手法は英語の迷惑メールにおけるそれと異なることが考えられる．本論文ではこの点を実験により調べる．

4 実験

日本語の迷惑メールにおいて，迷惑メール本体に無関係な文章（Word Salad）を付加することにより統計的フィルタリングの回避を行った場合の効果について実験を行った．

実験では，同一の短い迷惑メール本体（実際の迷惑メールをもとにしたもので，本文は日本語約 25 字，URL およびメールアドレス各 1 件からなる）に 100 通りの Word Salad を付加して迷惑メールを作成し，統計的フィルタリングによって判定を行い，フィルタが計算した迷惑メール確率を測定した．Word Salad として使用した文章は Yahoo!ニュース [15] からランダムに取得したニュース記事をもとに，全角文字のみを 50 字ずつに分割して 1 つの単位とし，迷惑メールに対して 1 個ないし複数個を付加した．また，迷惑メール本体がより長い場合（本文は日本語約 475 字，URL およびメールアドレス各 1 件からなる）についても比較のため実験を行った．

実験ではナイーブベイズ（ベイジアンフィルタリング）を用いた実装のひとつである Bsfilter[11] を使用し，迷惑メール確率の計算式には既定値の Robinson-Fisher 方式（閾値の既定値は 0.95）

表 1: Word Salad による bsfilter への影響 (元の迷惑メールが短い場合)

付加量	結果（回）				
（文字）	$p < 0.5$	$0.5 \leq p < 0.7$	$0.7 \leq p < 0.9$	$0.9 \leq p < 0.95$	$0.95 \leq p$
50	0	12	5	6	77
100	0	24	8	4	64
150	0	34	14	5	47
200	0	52	10	3	35
250	0	61	12	2	25
300	0	78	5	2	15
350	0	78	9	1	12
400	0	84	8	1	7

・Word Salad がない状態では，元の迷惑メールの迷惑メール確率 p は 1.0.

表 2: Word Salad による bsfilter への影響 (元の迷惑メールが長い場合)

付加量	結果（回）				
（文字）	$p < 0.5$	$0.5 \leq p < 0.7$	$0.7 \leq p < 0.9$	$0.9 \leq p < 0.95$	$0.95 \leq p$
100	0	0	0	0	100
200	0	2	0	1	97
300	0	12	11	2	75
400	0	19	10	4	67
500	0	35	12	4	49
600	0	41	11	4	44
700	0	53	9	3	35

・Word Salad がない状態では，元の迷惑メールの迷惑メール確率 p は 1.0.

を使用した．判定に用いる学習データは，正当な電子メールと迷惑メールを各 400 通学習したものを使用した．

5 結果

結果を表 1,2 に示す．Word Salad として元の迷惑メールに付加した文字数が増加するにつれて，おおむね迷惑メール確率が低下する傾向がみられた．また，元の迷惑メールが長い場合も Word Salad の効果は見られたが，同等の効果を得るためにはより多くの Word Salad が必要となった．

6 まとめ

本論文では，迷惑メールにランダムな単語や文章を付加することにより統計的フィルタリングを回避する攻撃について，ベイジアンフィルタを用いた実験を行い，その影響を計測した．実験の結果，日本語の迷惑メールにおいても Word Salad 攻撃が統計的フィルタリングによる迷惑メールの検出に影響を与えうることがわかった．

謝辞

本研究の一部は，財団法人セコム科学技術振興財団平成 15 年度研究助成「インターネット妨害障害に対する暗号論的対策技術の研究」の支援を受けている．

参考文献

[1] Sender Policy Framework (SPF), http://spf.pobox.com/.

[2] E. Gabber, M. Jakobsson, Y. Matias and A. Mayer, Curbing Junk E-Mail via secure Classification, *Financial Cryptography '98* , Anguilla, British West Indies, International Financial Cryptography Association, LNCS 1465, Springer, pp. 198–213 (1998).

[3] Penny Black Project, http://research.microsoft.com/research/sv/PennyBlack/.

[4] C. Dwork and M. Naor, Pricing via Processing or Combatting Junk Mail, *Crypto '92,* pp.139-147, 1993.

[5] M. Jakobsson, J. Linn and J. Algesheimer, How to Protect Against a Militant Spammer, *Cryptology ePrint archive,* report 2003/071, 2003.

[6] P. Graham, A Plan for Spam, http://paulgraham.com/spam.html.

[7] Procmail, http://www.procmail.org/.

[8] J. Graham-Cumming, How to beat an adaptive spam filter, *Spam Conference 2004,* 2004, http://www.jgc.org/SpamConference011604.pps.

[9] G. Wittel and S. Wu, On Attacking Statistical Spam Filters, *The first Conference on Email and Anti-Spam,* Mountain View, California, USA, 2004.

[10] Mozilla 1.3 Release Notes, modified February 2004, http://www.mozilla.org/releases/mozilla1.3/.

[11] Bsfilter, http://www.h2.dion.ne.jp/~nabeken/bsfilter/.

[12] Scbayes, http://www.shiro.dreamhost.com/scheme/wiliki/wiliki.cgi?Gauche%3ASpamFilter&l=jp.

[13] Popfile, http://popfile.sourceforge.net/.

[14] KAKASI, http://kakasi.namazu.org/.

[15] Yahoo!ニュース, http://headlines.yahoo.co.jp/hl.