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

2003－ＱＡＩ－8　(10)
2003／7／7

コンテンツに基づくマスメイルフィルタリング

和田俊和†　齋藤彰一†　泉裕†　上原哲太郎‡

†和歌山大学システム工学部／システム情報学センター
〒640-8510　和歌山市　栄谷　930番地
‡京都大学大学院工学研究科
〒606-8501 京都市　左京区　吉田本町

E-mail: {twada, shoichi}@sys.wakayama-u.ac.jp, yutaka@center.wakayama-u.ac.jp,
tetsu@info.kogaku.kyoto-u.ac.jp

あらまし　近年、マスメイル送信は大きな社会問題となっており、我々はこの問題に対処するためのシステムを構築している。Open Relay MTAの数に比べてマスメイルコンテンツの種類の方が圧倒的に少ないという事実に基づき、このシステムではコンテンツに基づくフィルタリング方式を採用している。これまで、受信メイルに含まれる単語ヒストグラムやn-グラムに基づいて識別を行うものが多く実装されているが、同種のメイルが多数送受信されるというマスメイルの特性に着目すると、コンテンツDBを構築して受信メイルがDB内のエントリに一致するか否かで対処することが有効である。本研究では、マスメイルを自動的に収集し、そのチェックサムをDNSに登録してフィルタリングを行う。すでに、マスメイル収集、チェックサム計算、およびDNSへの登録、が行われており、これを参照したメイルフィルタも稼動している。本稿では、このシステムの概要、運用実績と今後の展望について述べるとともに、マスメイル収集について協力を呼びかける。

キーワード マスメイル、SPAMメイル、不正中継MTA、コンテンツベースフィルタリング、

Contents based Mass-Mail Filtering

Toshikazu WADA†,　Shoichi SAITO†, Yutaka IZUMI†　and Tetsutaro UEHARA‡

† Faculty of Systems Engineering, Wakayama University
Sakaedani 930, Wakayama-shi, Wakayama, 640-8510 Japan
‡ Graduate School of Engineering, Kyoto University
Yoshida Hon Machi, Sakyo-ku, Kyoto, 606-8501 Japan

E-mail: {twada, shoichi}@sys.wakayama-u.ac.jp,　{izumi, tetsu}@center.wakayama-u.ac.jp

Abstract Recently mass-mail sending is becoming a serious problem that prevents normal commerce and advertisement on the Internet. We are now developing a system solving this problem. Existing mass-mail filtering systems can be classified into two types; 1) open relay MTA based filtering systems and 2) recognition based contents filtering systems. Since the number of open relay MTA is much bigger than the number of mass-mail contents, contents based mass-mail filtering is much effective than the open relay based filtering. However, the recognition based contents filtering requires time consuming training and tuning. Unlike these approaches, our system consists of two parts; mass-mail checksum database and clients. The checksums are delivered via the DNS and the clients. The prototype database system is currently running, i.e., mass-mail contents sent to some dummy addresses are analyzed, normalized, and their checksums are stored into the DNS. As well, some clients referring the DNS are developed. This report presents the overview of our current system and the future extensions.

Key words　Mass-mail sending, SPAM mail, Open relay MTA, Contents based Filtering

1. はじめに

SPAM メイルに代表される、いわゆる迷惑メイル（unsolicited e-mail）は、公序良俗に反するサイトへの誘導、消費者トラブルの誘発、メイルトラフィックの増大等、様々な問題を引き起こしており、世界的な規模で様々な対応策が検討されている。また、メイル拡散型ワームやウイルスなども、正常な電子メイルの利用を妨げるものとして、その防御法がこれまで多数開発されてきた。

最初に用いられた手法は、Open Relay MTA の IP アドレスを記録した DNS を参照して、Open Relay ホストからのメイルを受け取らないようにするシステムである[1][15]。しかし、このシステムには以下のような欠点がある。

1）世界中に 100 万台以上存在すると言われる Open Relay MTA を全て網羅したデータベースを構築することは不可能である。
2）マスメイル送信業者が、この DNS を参照してマスメイル送信に利用している。
3）一旦 Open Relay MTA としてデータベースに登録されてしまうと、そのホスト経由で送られてきたメイルは、正規の利用者からのメイルも含め全てブロックされてしまう。
4）基本的に MTA 単位でしか利用できずエンドユーザ単位では利用が困難である。

このような問題点があるため、最近ではメイルのコンテンツに基づいて、マスメイルフィルタリングを行う方法が盛んに検討され、実装も進められている([16]- [19]などは主にメイルボディ部のコンテンツに着目したものであり、ヘッダ情報とボディ部の情報を組み合わせて識別するもの[20]もある。)。これらのシステムの多くは、メイルに含まれるキーワードや n-gram などの統計的性質に基づいて未知のメイルがマスメイルであるか否かを識別するタイプのものである。この種のシステムは、個人単位で利用することができ、しかも個別にチューニングできるという特長がある反面、識別の精度と速度などが問題となる。また、これらの識別システムが公開されれば、その識別アルゴリズムを回避する方法が調べられ得るため、高い識別率が維持できるとは保証できない。

本稿でも、メイルコンテンツに基づくマスメイルフィルタリングの方法を検討するが、我々は、個々のメイルをコンテンツによって識別する問題そのものは主題として取り扱わない。むしろ、これらの方法、あるいは人手によって発見されたマスメイルコンテンツのチェックサムを計算して公開し、これを参照することによって、マスメイルを排除する方式について検討を行う。この方式は、

1）マスメイルは基本的に同じ内容のメイルが大量に送受信されるため、同種メイルの識別を何度も繰り返す必要はない。
2）同じデータベースを参照している集団には、登録されたメイルが届かないようにすることができる。
3）クライアントソフトウエアは MTA や POP・IMAP サーバだけでなく、POP・IMAP クライアントやメイル整理ツールまで、幅広く対応でき、包括的なマスメイルフィルタリングが実現できる。
4）一度データベースに登録されたものが抹消されることは間違いを除いてありえない。そのため、ローカルキャッシュを参照することができ、データベースの負荷を必要以上に上げないアプリケーションの構築が行える。

等の利点がある。特に、1）2）は集団的なマスメイル駆除を行うことのメリットであり、3）もまた個別の識別システムでは実現することが困難な機能である。また、このデータベースの作成過程でマスメイルコンテンツ自体を収集しているため、識別型フィルタの学習データも副産物として得られる。

BT の顧客に対する SPAM フィルタリングや、AOL の新クライアントに Anti-SPAM 機能が盛り込まれるなど同種のマスメイルフィルタリング方式はすでに一部実用化されているが、日本語テキストへの対応や、特定組織内での利用など制限が多いためまだ十分には広まっていない。

このようなデータベースを収集するには、マスメイルを効率よく集めてこなければならないがその方法は必ずしも明確ではないという問題がある。

また、個々のマスメイルのボディ部分にも送信相手に固有の文字列や無意味な文字列が埋め込まれているケースが多く、単純なチェックサム計算では同種のマスメイルに対して異なるチェックサムしか得られず問題となる。

さらに、チェックサムによってマスメイルと判断された場合に、いきなりメイルを削除もしくは受け取らない、などの二者択一的アクションを行わず、誤ってマスメイルと判定される正常なメイ

ルがあり得る事を考慮したユーザインタフェースを提供する必要がある。

以下、これらの問題に対して検討を行った結果について述べ、続いてシステムの概要について述べる。

2. マスメイルデータベース構築における課題

マスメイルデータベース構築に当たって、必要になることは、以下のように整理できる。

- できる限り多種のマスメイルコンテンツを、流通しはじめてからすぐに収集し、データベースに反映させること。
- 収集したメイルの内容のうち、送り先に固有の情報や解析を避けるために埋め込まれたランダムな文字列を回避して、マスメイル毎に固有のチェックサムを求める方法の開発。
- マスメイルデータベースを参照してメイルをブロックするクライアントの開発

以下、上記の内容について説明する。

2.1 マスメイルの収集

マスメイルの収集は、現在以下のようにして行っている。

- 現在マスメイルを受信しており、実質的な使用者が存在しないメイルアドレスから転送を行う。
- 囮メイルアドレスを WEB ページのコメントもしくは、実質的にクリックできない mailto アンカーに埋め込む。
- 囮メイルアドレスに関して、マスメイル送信業者にメイル送信を行わない旨のメイルを送る。
- 新規ドメインを取得し、webmaster@domain, info@domain などそのドメイン内の全てのメイルアドレス宛のメイルを収集する。
- 当該プロジェクト協力者宛のメイルを分析対象とする。

上記の方法で、メイル（マスメイル候補）を収集した後、以下のようにしてメイルコンテンツの分析および登録を行う。

- コンテンツチェックサムデータベースにすでに登録されているか否かをチェックする。
- 登録されていれば、すでに登録されている内容へのメイルとしてデータベースに格納する。
- 登録されていなければ、メイルコンテンツの識別を行い、マスメイルと判定された場合には新規マスメイルコンテンツとしてデータベースに登録を行う。
- マスメイルと判定されなかった場合は、メイルスプールに書き込む。

という順序で登録処理を行う。メイルコンテンツの識別方法については、ヘッダ部分に着目して分析を行う方法と、メイルボディ部分を解析して識別する方法の２つが考えられるが、現時点では、メイルボディの識別は開発中であり、ヘッダ情報（中継 MTA の IP アドレス情報）に基づいて、識別を行う方法を実装している。

2.2 チェックサムデータベースの構築

マスメイルのボディ部分は、下記の理由により、同じコンテンツであっても異なる文字列となっていることが一般的である。

1. base64, quoted-printable などのエンコーディングがかかっているケースがある。
2. ユーザ毎に割り当てられた特殊な ID 文字列や、ユーザのメイルアドレス、あるいは HTML のコメントタグなどが挿入されているケースがある。

このうち、1.に関しては、デコードをして対処する。2.に関しては、正規文法として受理すべき文字列を決め、下記のヒューリスティクスを使い、メイルコンテンツから不要な文字列を削除している[1]。

- 英数字の混じった文字列
- コントロールコード
- アンカーURL 以外の HTML タグ
- メイルヘッダの Delivered-To フィールドに現れる文字列

上記の正規化を行った後に MD5 によってチェックサムの計算を行っている。

MD5 は異なるデータに対して同じチェックサムが得られる可能性が極めて低く、実質上確率０と見なしてよいこと、および、元の正規化された文字列が復元できない、という条件が満たされているため、当システムのチェックサム計算法としては最も妥当なものであると考える。

このようにして得られたマスメイルのチェックサムデータは DNS に自動的に登録され、公開されている。

2.3 マスメイルの取り扱い

---

1 チェックサム計算の詳細アルゴリズムは、これを回避する措置が取られないよう、詳細については公開できない。

上記データベースを参照して動作するクライアントプログラムとしては、多数のものが考えられるが、現在は単純なメイルフィルタ、POP サーバ、POP クライアント、などを開発している。

メイルフィルタ： メイルチェックサムを計算して DNS に対する問い合わせを行う。その結果に従い、メイルの通過とブロックを行う。この実装の欠点は、ブロックしたメイルに正常なメイルが含まれていた場合に対処することができないということである。

POP サーバ： POP サーバでチェックサムを計算し、DNS を参照してマスメイルと判定された場合は、マスメイルである旨を表現するメイルヘッダを挿入する。メイルリーダの側でこのメイルヘッダに基づいて自動分類するようにしておけば、仮に誤って正常なメイルがマスメイルと判定されてしまっても、ユーザはそのメイルを失うことはない。

POP クライアント： 上記 POP サーバで行うのと同様の内容を POP クライアント側で行う。

これらいずれの場合も、不正中継 MTA の IP アドレスが Received From 行に現れた場合にマスメイルとして検出するという機能も追加できる。また、後者のチェックによってマスメイルが検出された場合、DNS への自動登録を行うということも実現できる。このように、クライアントプログラムにデータベースへの自動登録機能を組み込み、それが安全に運用できれば、結果的にチェックサムの参照によるマスメイル検出の確率が増加すると考えられる。

3. 運用実績

当該システムは、昨年秋からプロトタイプシステムの運用を開始し、1 月 16 日から正式運用を開始した。以下、これまでの運用記録から、その一部を抜粋して紹介する。

3.1 データベース登録数

図 1 に新規マスメイルおよび既知マスメイルとして検出されたメイルの件数を表すグラフを、図 2 にこの結果を累積したグラフを、それぞれ示す。

運用開始直後から新規マスメイルコンテンツは一日平均 15 通程度が蓄積されていたが、4 月ごろになると自然増によって 20-25 通程度に増加している。また、4 月末に

- 囮メイルアドレスに関して、マスメイル送信業者に対してメイル送信を行わない旨伝えるメイルを送る。

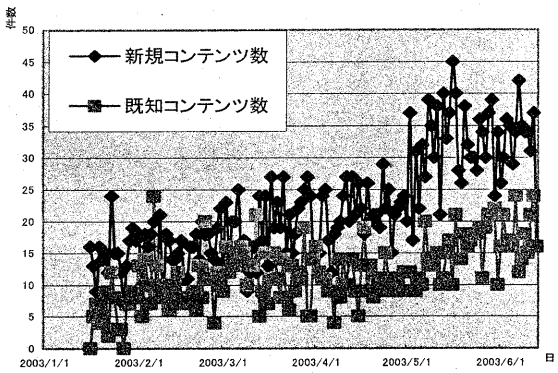

図1：新規・既知コンテンツの推移

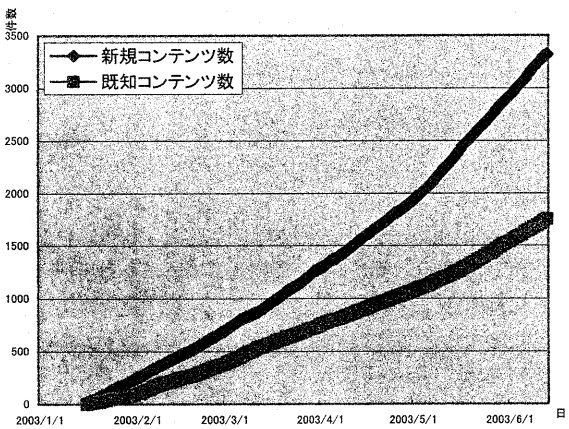

図2：新規・既知コンテンツの累積件数

という作業を行った約 1 ヵ月後頃から 1 日平均 35 通程度の新規マスメイルが収集できるようになっている。

また、先にデータベースに登録された同じ内容のメイルを受け取った際に、既知のメイルとして検出された件数も同じグラフに示す。このグラフから、既知のメイルは新規メイルの約半数であることが分かる。

すべての新規マスメイルに対して既知メイルが届くという状況にならなければ、十分なマスメイルソースが確保できたとは言えないが、現状では既知メイルの検出件数が新規マスメイルの件数を越えていないことからも分かるように十分な数のマスメイルが集まっているとは言えない。このことをより詳しく調べた結果を図 3,4 のグラフに示す。

これらのグラフは、1 回だけしか届かなかったメイルと、複数回届き既知メイルとして検出されたメイルコンテンツの数を表しており、図 3 はこれらの数の推移、図 4 はその結果を累積したものである。この結果から、1 回だけマスメイルとして検出されたコンテンツに比べて、複数回検出されているのは、

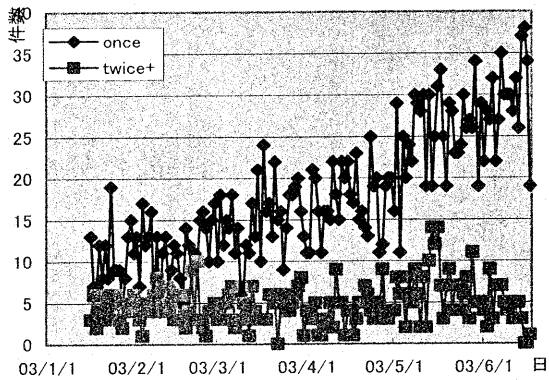

図3:単数・複数回検出のメイルコンテンツ数の推移

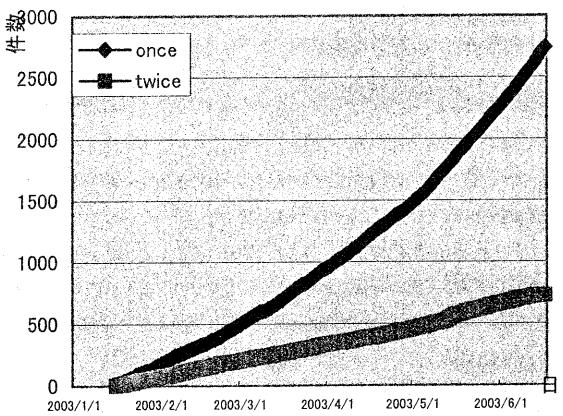

図4:単数・複数回検出のメイルコンテンツ数の推移(累積)

約 1／4 程度しかない。

この原因としては、

- 十分な数のマスメイルが集まっていない
- 同じコンテンツに対して異なるチェックサム値が計算されている。

という 2 つの可能性が考えられるが、調査の結果、後者のケースは稀であり、マスメイルの種類の多さに比べて、十分なマスメイルソースが押さえられておらず、かろうじて収集しただけのマスメイルが多いことが主な原因である。

3 月の時点ではこの割合が約 1／2、5 月には 1／3 であったのに対し、現在では 1／4 と低下している。これは、比較的少ない種類のマスメイルが大量に出回っていた状態から、より多くの種類のメイルが送られるようになったことを意味しているものと思われる。この割合は、より多くのマスメイルソースを確保し、大量のマスメイルを収集することで改善できるものと考えられる。

一方、これまでのマスメイルの到着時刻を時間別に示すと、図 5 のようになり、ほぼ時間的に偏りなく均一にマスメイルが到着していることが分かる。

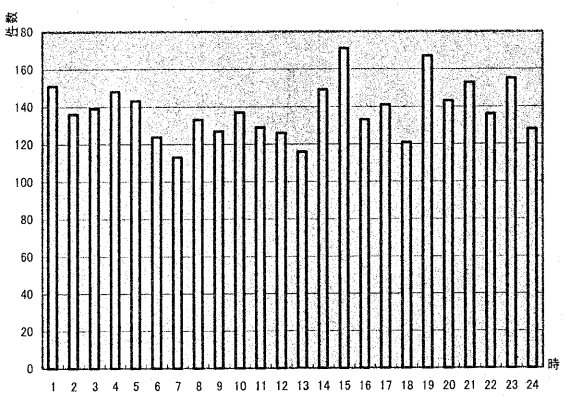

図5:時間別マスメイル到着件数

新規マスメイルが登録されてから、同じ内容のマスメイルを受け取るまでの時間は、マスメイルの時間的特性を調べる意味で重要なデータとなる。すなわち、計算されたチェックサムをどの程度 DNS 上に保持しておくべきか、などを決定する際に役立てることができる。図 6 は、複数回検出されたマスメイルコンテンツについて、同一のマスメイルが再び検出されるまでの日数を表している。

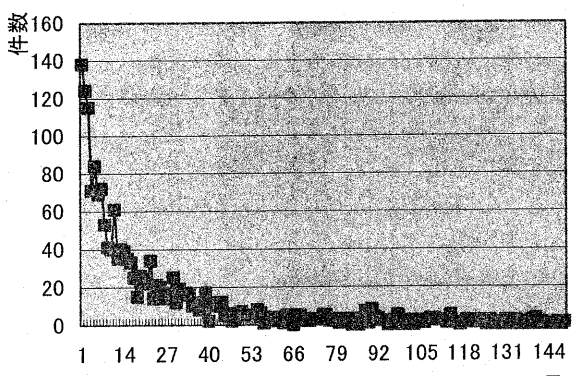

図6:同一マスメイル到着までの日数

このグラフによると、ほぼ 2 ヶ月以上経つと同じ内容のメイルが届く可能性は極めて低くなると言え

る。また、当然の結果ではあるが、マスメイルを受信した当日に同じマスメイルが届く可能性は極めて高く、できるだけ最初に送られたマスメイルをキャプチャすることが、フィルタリングを行う上で極めて重要であるといえる。

4. 考察と展望

マスメイルコンテンツのデータベースを作成し、同種のメイルをブロックすることは、マスメイルに対する集団的防御基盤を構築することを意味している。具体的には、識別型フィルタなどで個々人がマスメイルに対処しているという現状に、個々人がマスメイルとしてブロックしたメイルをデータベースに登録し、それを他者が利用して同じメイルをブロックすることにより、集団としてのマスメイルの防御能力を高めることを指す。このように集団としてのマスメイルフィルタリングが軌道に乗れば、マスメイル送信そのものを無力化することに役立つはずである。

現状のマスメイルコンテンツデータベースは、3章で述べたようにまだまだ十分な数のコンテンツが蓄積されているわけではなく、さらに多くのマスメイルを収集する必要がある。より多くのマスメイルコンテンツを収集するために、協力頂ける組織・個人は是非ともご連絡いただきたい。このようなマスメイル収集が軌道に乗れば、より確実なマスメイルフィルタリングを行うことができるようになるはずである。

5. まとめ

本稿では、現在我々が検討を進めているマスメイルデータベースの概要および、運用実績について述べた。今後は、これまでの成果を踏まえ、今後は、幾つかの組織およびプロバイダでの試験運用を開始し、さらに多くのマスメイルコンテンツを収集する予定である。また、実際のマスメイルによって学習された識別型のフィルタによるマスメイル検出機構を追加するとともに、各クライアントソフトウエアの充実を図る予定である。

文　献

[1] http://www.ordb.org/
[2] http://www.mail-abuse.org/rbl/
[3] http://relays.osirusoft.com/
[4] http://spews.org/
[5] http://selwerd.cx/xbl/
[6] http://www.spamcop.net/
[7] http://fabel.dk/relay/
[8] http://rbl.v6net.org/
[9] http://www.five-ten-sg.com/
[10] http://basic.wirehub.nl/spamstats.html
[11] http://www.blars.org/block.html
[12] http://www.rfc-ignorant.org/
[13] http://www.leadmon.net/spamguard/
[14] http://dsbl.org/
[15] http://www.spamhaus.org/
[16] http://www.paulgraham.com/spam.html
[17] http://www.mozilla.org/mailnews/spam-howto.html
[18] http://www.mcafee.com/myapps/msk/default.asp
[19] http://www.symantec.com/sabu/nis/nis_pr/index.html
[20] http://spamassassin.rediris.es/index.html