2004－DSM－32　(13)
2004／3／29

マスメイルデータベースと
それを用いたマスメイル検出システム

松浦広明†　齋藤彰一†　上原哲太郎††　泉裕‡　和田俊和†
†和歌山大学システム工学部　††京都大学大学院工学研究科
†和歌山大学システム情報学センター

概要　マスメイルによる被害は日に日に増大している．本研究では，マスメイルを検出するために，マスメイルデータベースと POP サーバからなるシステムを構築した．データベースでは，マスメイルを自動的に収集し，各マスメイルを正規化・解析したのち計算されたチェックサムを蓄積する．POP サーバでは，転送するメイルのチェックサムを同様に計算し，データベースに問い合わせることで，マスメイルかどうかを判断する．このシステムでは，チューニングや学習は不要である．処理速度が速く，データベースの内容が正しい限り誤検出は起きないのが大きな特徴である．このシステムを和歌山大学内の POP サーバで稼動させた結果，マスメイルの認識率はおよそ 80-90%であった．

Mass-mail Database and Mass-mail Detection System.

Hiroaki Matsuura† Shoichi Saito† Tetsutaro Uehara††
Yutaka Izumi‡ Toshikazu Wada†
†Faculty of Systems Engineering, Wakayama University
††Graduate School of Engineering, Kyoto University
†Center for Information Science, Wakayama University

Abstract Problems caused by mass-mail have been more and more serious. We have constructed a system for detection of mass-mails, which is composed of a mass-mail database and a POP server. The database collects mass-mails automatically, and registers checksums calculated by analyzing the contents of the each mail. The POP server computes checksums of the handling mail and asks the database whether the checksums have been registered as mass-mail. This system requires no efforts of tuning and learning, runs fast, and does not mistake in mass-mail recognition as long as the database is valid. The server has run on a POP server in Wakayama University. Almost 80-90% of mass-mails have been detected successfully.

1. はじめに

不特定多数に大量に送信されるメイル（マスメイル）は，受信者へ大きな負担，公序良俗に反する内容，メイルトラフィックの増大，ウイルスメイルの拡散など，多くの問題を世界的な規模で引き起こしている．日に日に増え続けるマスメイルを一刻も早く止めるために，様々な対応策が検討されている．

マスメイルを判別する場合，コンテンツ同士を単純に比較する方法を用いることはできない．なぜなら，マスメイルの送信者は，html の仕様を利用してコンテンツの表記方法を変化させたり，本文とは関係のない文字列をコンテンツに加えたりするなどの方法で，「意味は同じだが違う」コンテンツを大量に作成し，送信するためである**[1]**.

また，コンテンツのすべてではなく任意のキーワードを抽出･検索する方法がある．この方法は考え出された当初は効果的であったが，今では有用ではなくなった．マスメイル送信者は，キーワードの綴りをわざと書き換えて，検出されないようにしているためである．

以上のような判別の困難さを解決するため，メイルのコンテンツに含まれる各トークンの統計的性質に基づいてマスメイルの識別を行う方法が盛んに研究されている．中でも多いのが Bayesian Filter を利用した種々のメイルフィルタリングツールである**[2]-[6]**．これらのツールは，個人単位で利用することができ，チューニングできるという特長がある．その反面，学習させるまでに時間がかかること，学習のデータによって識別の精度が左右されること，処理に時間がかかること，日本語のような分かち書きされていないテキストは不得意であること，などの不具合がある．また，最近のマスメイル送信者は，一般的なビジネス文書やニュース記事を含むようなマスメイルを送ることによって Bayesian Filter に誤判定・誤学習を起こさせるなど，マスメイルが検出されないよう工夫しつつある．

マスメイル検出システムは，処理が速く，誤検出がなく，ユーザが手軽に使えるようなシステムであることが望ましい．本研究では，マスメイルを蓄積するデータベースを活用して，マスメイルの判別を行う．

2. マスメイル判別システム

2-1 概要

本研究のシステムは，マスメイルを蓄積するサーバと，それを活用する POP サーバの二つのサーバからなる．

マスメイルデータベース（図１）では，マスメイルを自動的に収集する．集められたマスメイルに対して，コンテンツのランダム要素や非表示文字を排除する「テキストの正規化」および「URL の抽出」を行う．得られたそれぞれのデータから MD5 法によりチェックサムを計算し，データベースに蓄える．データベースは，リモートからの利用が容易な DNS を用いる．

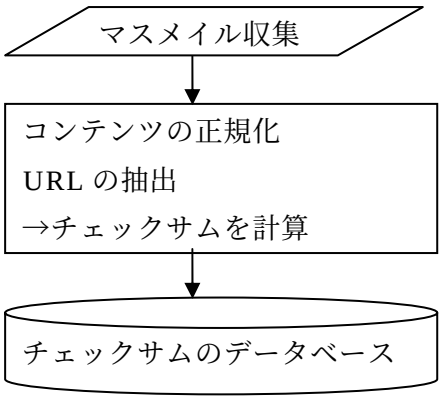

図１　マスメイルデータベースの概要

POP サーバ（図２）では，処理中のメイルに対してデータベースと同様の処理でチェックサムを計算し，データベースに問い合わせる．チェックサムがデータベースにあれば，処理中のメイルをマスメイルだとみなし，メールヘッダにキーワードとして"X-MMS-RESULT: 1"を挿入する．

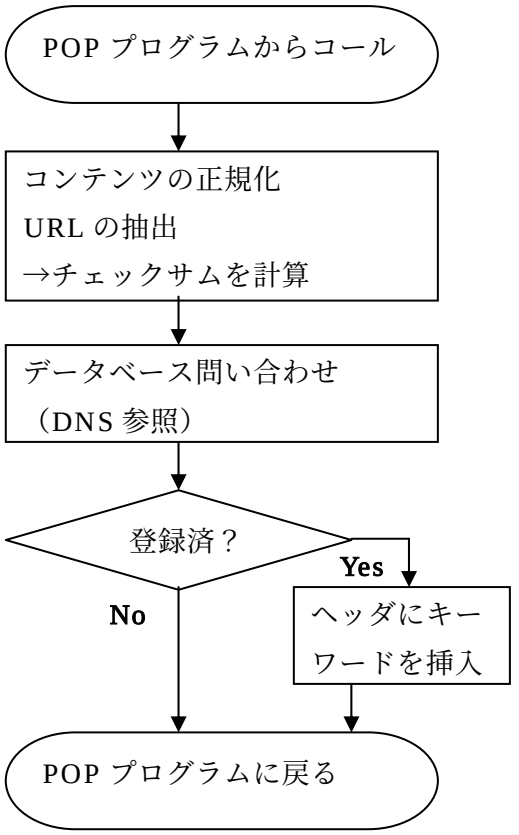

図２　POP サーバの概要

2-2 本システムの特徴

本システムでは，次のような利点が挙げられる．

１） データベースが正常である限り，誤認識は起こらない．

２） 認識器のような複雑な計算が必要なく，高速に動作する．

３） ユーザはメールリーダで「メールヘッダ中のキーワードによる，フォルダ振り分け」を設定するだけでよい．

４） チェックサム計算の部分をライブラリ化しているため，procmail のようなメイルフィルタと組み合わせて実行したり，メイル整理ツールを作成したりと，包括的なマスメイルフィルタリングができる．

3 マスメイルデータベース

3-1 マスメイルの収集

現在行っているマスメイルの収集方法を以下に述べる．

・マスメイルを受信するためだけに用意したメイルアドレス（おとりアドレス）から転送を行う．

・マスメイルを受けたいメイルアドレスを 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．

・マスメイルに「送信を止めるためにはメイルを送るように」など指定されている場合は，マスメイル送信者にメイルを送る（これによって，マスメイルの送信者はそのメイルアドレスが有効であることを知る）．

・新規ドメインを取得し，そのドメイン宛てのすべてのメイルをマスメイルとして収集する．

3-2 データベースの状況

集められたマスメイルの累積登録数（図３），一日あたりの登録数（図４），および時間帯あたりの登録数（図５）を示す．現在 4 万通弱のマスメイルが収集されており，一日あたりの登録数は増加傾向にある．また，時間帯による変化はほとんどないことがわかる．

図３　累積メイル数

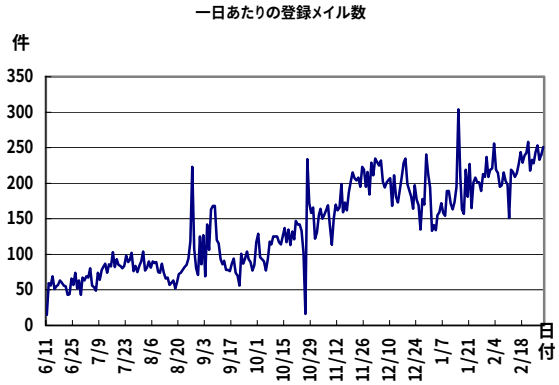

図４　一日あたりの登録メイル数

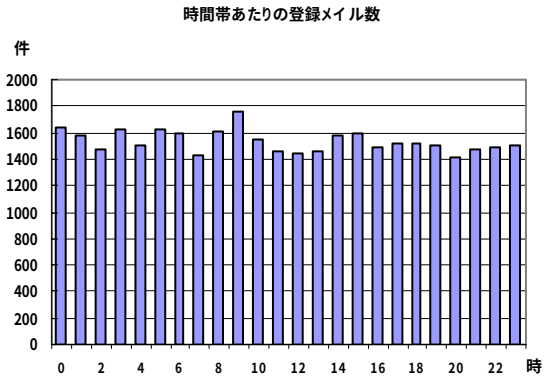

図５　時間帯あたりの登録メイル数

4. チェックサムの計算

前述したように，データベースおよび POP サーバでは，「テキストの正規化」および「URL の抽出」を行い，それぞれの結果に対して，チェックサムを計算する．

「URL の抽出」において，本来マスメイルとは関係ない URL がマスメイルに書かれている場合，誤ってその URL をデータベースに登録してしまう危険性がある．そこで，「ホワイトリスト」を外部ファイルに定義して，マスメイルとは無縁の URL を保護している．

この一連の処理を行うルーチンはライブラリ化してあるため，容易に POP サーバやメイルフィルタリングツールに組み込むことができる．現在，ipop3d，qmail-pop3 に組み込んで，動作を確認している．

チェックサム計算にかかる時間は 3 万通のメイルでおよそ 80 秒であり，一通あたり 0.00263 秒であった（CPU: Intel Pentium 4 2.4GHz，OS: Red Hat Linux 7.3，Memory: 1GByte）．なお，アルゴリズムの詳細は，マスメイル送信者に本システムを回避する情報を与えかねないため公開できない．

5. POP サーバの運用

5-1 認識率

4.で述べたライブラリを ipop3d に組み込み，和歌山大学システム工学部にて運用している．ユーザにより違いはあるが，マスメイルのおよそ 80～90%を正しく認識している．しかしまだ 10～20%のマスメイルを認識していないことから，データベースの拡充や正規化アルゴリズムの改善などが必要である．

一方，正常なメイルをマスメイルと誤認識することはない．運用当初はデータベースに正常なメイルが紛れ込んでいたことや，ホワイトリストが不十分であったため数件生じたが，現在では皆無である．

5-2 メイルトラフィックの状況

一日における全 POP 数と検出マスメイル数（図６），全 POP 数に対する検出マスメイル数の割合（図７）を示す．平日・休日の周期がみられ，検出率はおよそ 6〜12%を推移している．

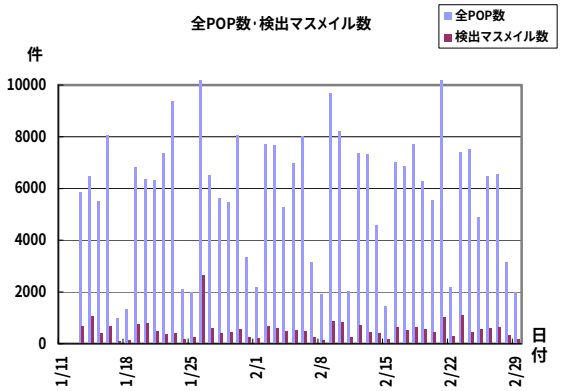

図６　全 POP 数と検出マスメイル数

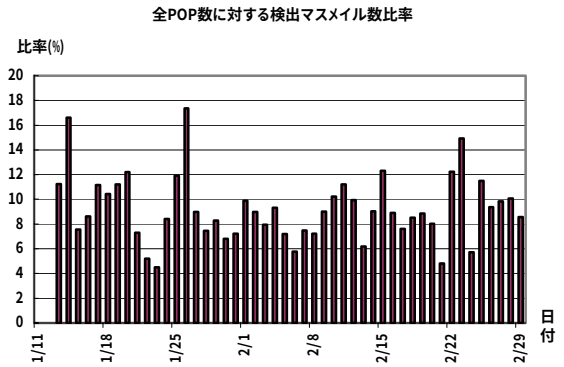

図７　全 POP 数に対する
検出マスメイル SPAM 数比率

5-3 マスメイルの検出時に用いられたチェックサムの種類

一日におけるマスメイルの検出が，正規化したコンテンツのチェックサム（BODY）でなされたのか，抽出した URL のチェックサム（URL）でなされたのかを検討した．

全体の検出マスメイル数に対する，BODY での検出率と URL での検出率を，図８に示す（一つのマスメイルが BODY と URL の両方で検出される場合もあるため，BODY の比率と URL の比率との和が 100%以上になることがある）．この結果から，URL による検出が BODY による検出のおよそ 2〜3 倍ほど高い比率を示していることが分かった．URL はマスメイルの判別に大きな手がかりとなっていると言える．BODY の検出率が低いのは，コンテンツは URL よりもランダム要素が大きく，正規化の段階でランダム要素を完全に除去できなかった理由もあると考えられる．

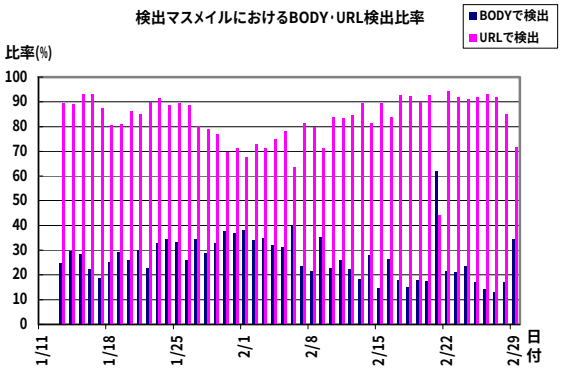

図８　検出マスメイルにおける
BODY・URL 検出比率

5-4 チェックサム登録日時とマスメイル検出日時との時差

POP サーバが任意のメイルをマスメイルと判断した日時と，判断の理由となったチェックサムがマスメイルデータベースに登録された日時を比較した．これによってチェックサム登録からマスメイル検出までの時間を知ることができる．ただし，データベースには同じチェックサムを持つマスメイルが一つ以上登録されているため，求める解は複数ある．解のうちの最短の時差を求めた．その結果を図９に示す．

これによると，POP サーバでマスメイルだと判断されたチェックサムのうち，50%

超が 24 時間以内に届いたマスメイルから登録されたチェックサムであり，また，ほとんどすべてが 10 日以内に届いたマスメイルから登録されたチェックサムである．このことから，POP サーバで検出されるマスメイルとマスメイルデータベースに届くマスメイルは，ほぼ同時期に届けられていることがわかる．より多くのマスメイルをより早くデータベースに登録すれば検出率の向上が見込まれる．

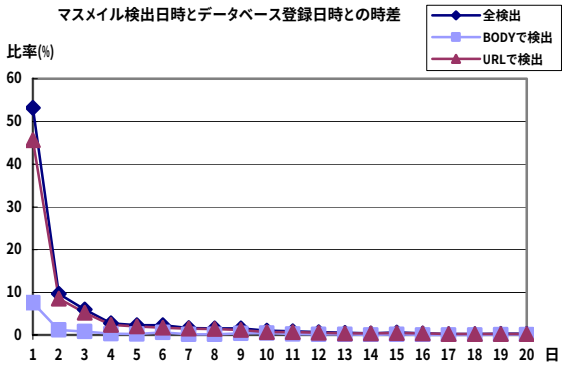

図９　マスメイル検出日時と
データベース登録日時との時差

6. まとめと今後の展開

自動的に収集したマスメイルを正規化・解析してチェックサムを計算し，そのチェックサムデータを配信するマスメイルデータベースと，それを参照してマスメイルを検出する POP サーバとを用いて，マスメイルを検出するシステムを構築した．高速性，正確性，簡便性などが秀でている．

現在よりマスメイル検出率を上げるために次の対策を行っている．

1） ランダム要素を取り除くアルゴリズムの性能を上げる．

2） 検出率を上げるためには数多くのマスメイルを集める必要がある．spam archive**[7]**のマスメイルデータベースを利用することを検討中である．

3） 現アルゴリズムでは，「正規化したコンテンツ」と「抽出した URL」でチェックサムを計算している．メイルアドレスや電話番号をも抽出する新アルゴリズムをテスト中である．現行版よりも高い検出率を得ている．

これらを反映させた新システムは，今年度中に運用を開始する．

本システムは和歌山大学システム工学部で運用中であるが，来年度は，和歌山大学・京都大学・大阪大学など各研究機関のメイル管理者の協力を仰ぎ，運用を拡大してゆく予定である．

参考文献

[1] John Graham-Cumming, "The Spammer's Compendium", 2003 SPAM CONFERENCE (http://sourceforge.net/docman/display_doc.php?docid=14869&group_id=63137).

[2] "A Plan for Spam", http://www.paulgraham.com/spam.html.

[3] "Getting Started with Mozilla Mail Spam Filtering", http://www.mozilla.org/mailnews/spam-howto.html.

[4] "McAfree Security", http://www.mcafee.com/myapps/msk/default.asp.

[5] "Spam Assasin", http://spamassassin.org/.

[6] "POP File", http://popfile.sourceforge.net/.

[7] "SpamArchive.org", http://www.spamarchive.org/.