情報漏洩を防ぐ Web メールフィルタリング手法の提案と設計

浦川順平 † 鈴木信雄 †† 鈴木健二 †
† 電気通信大学 ††KDDI 株式会社

1 はじめに

現在, 企業での情報漏洩が問題となっている. その情報漏洩はハッキング, ウィルスなどの外部要因ではなく, 内部の人間による情報の盗難, 流出などの要因が多い [1]. その内部要因の一つとして, メールを利用したものがある. このため, 多くの企業ではメールクライアントソフトを利用して発出されるメールの監視, 規制を行うシステムを導入している. しかしこれらのシステムでは Web メールを規制することはできない. また,Web フィルタリングシステムを導入している企業もあるが, このシステムは暴力・アダルトなどの有害サイトの閲覧を規制する目的で開発されており, Web メールサイトを適切に規制することはできない. そこで本稿では, 世の中に広く普及している Web メールサイト利用時のパケットの特徴を調査・検討し,Web メールの使用を検出・制限するシステム (図 1) の提案と設計を行ったので, 以下にその概要を説明する.

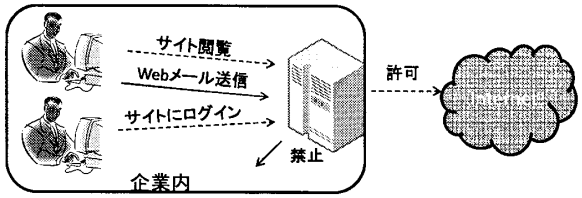

図 1: システム概要図

2 Web メールと既存のメールフィルタリング

Web メールを利用する際, ユーザが実際に通信を行うのはメールサーバではなく Web サーバであり, サーバとの通信には HTTP が使用されている。したがってサーバ・クライアント間のトラフィックの中から POP や SMTP のようなメールプロトコルに基づいてメールのトラフィックを検出するシステムでは Web メールを規制することはできない.

また,Web メールの使用を検出・制限するために既存の Web フィルタリングシステムを利用する方法があるが, このシステムはブラックリスト方式に大きく依存しているため, 未知のサイトを規制することができないという問題点がある. その問題点を補うために, 特定のキーワード (例.From, to など) の出現頻度により規制する方式も利用されていることがあるが, これは規制対象以外のサイトを規制してしまう可能性を含んでいる. そのため, このシステムでは,Web メールの使用を検出・制限するという目的を十分に達成出来ない.

Proposal and Designing of Web-Based Email Filtering for Information Leak Protection
Junpei Urakawa†, Nobuo Suzuki†† and Kenji Suzuki†
The University of Electro-Communications† and KDDI Corporation ††

3 Web メールの特徴

Web メールの使用を検出・制限するためには, Web メールサイトに接続し, メールを送信するという一連の処理の間にサーバ・クライアント間でやりとりされるパケットの特徴を知る必要がある. そこで, 世の中に広く普及している 43 種類の Web メールサイトにおいて, メール作成時にサーバから送信されるパケット, メール送信時にクライアントから送信されるパケットの特徴を調査した. 調査結果を以下に示す.

(1) クライアントが入力した情報はいずれの Web メールサイトにおいても, 全て POST メソッドによりサーバへ送信されていた. また, 図 2 に示すように, 多くのサイトでは POST メソッドが指定された form タグ内に,From, To, Cc, Bcc, Subject, Body, メールアドレスなどのメール特有のキーワードが出現していた.

```
<body>
<form method="post"~>
宛先 <input type="text" name="to"~~>
Cc <input type="text" name="cc"~~>
</form>
</body>
```

図 2: Web メール作成時の html 文書

(2) メール作成ページに推移する際, 整形された html 文書 (図 2) がサーバから送信されるものの他に,Web メールサイト接続時にサーバから送信された javascript によりクライアント側で html 文書を生成するものがあった. この場合, サーバから送信された javascript コードに, (1) で述べたような構造の特徴は現れていなかった. これは図 3 のように DOM(Document Object Model) を利用して html 文書を生成していたためだと考えられる.

```
var body = document.getElementsByTagName('body')[0];
var form = document.createElement('form');
form.method = 'post';
body.appendChild(form);
var text = document.createTextNode('宛先');
form.appendChild(text);
var to = document.createElement('input');
to.name = 'to';
form.appendChild(to);
```

図 3: html 文書を生成する javascript コードの例

(3) クライアントから送信されたパケットは Content-Type が application/x-www-form-urlencoded の場合は図 4, multipart/form-data の場合は図 5, html/xml の場合は図 6 のようになっており, それぞれ`&to=,name="to",<to>`のようにメール特有のキーワードが出現していた.

```
&to=urakawa@infnet.cs.uec.ac.jp&cc=&bcc=
&subject=kennmei&body=honnbunn
```

図 4: application/x-www-form-urlencoded

```
----------11538186919912
Content-Disposition: form-data; name="to"

urakawa@infnet.cs.uec.ac.jp
----------11538186919912
Content-Disposition: form-data; name="subject"

kennmei
----------11538186919912—
```

図 5: multipart/form-data

```
<to>
    <email>urakawa@infnet.cs.uec.ac.jp</email>
</to>
<subject>kennmei</subject>
<simplebody>
  <html>honnbunn</html>
  <text>honnbunn</text>
</simplebody>
```

図 6: html/xml

(4) 調査を行った内の 1 種類では, メール送信時に特殊な構造のパケットを送信していた. そのパケットのペイロード部は URL エンコードされており, それを URL デコードすると図 7 のような構造となっていた. しかし, その構造は規則性を持ったものであり, メール特有のキーワードも出現していた.

```
requests=[{"To": "urakawa@infnet.cs.uec.ac.jp,",
 "Cc": "", "Bcc": "", "Subject": "kennmei",
 "RichBody": "<font face=¥"Arial, Helvetica,
sans-serif¥">honnbunn</font>", "PlainBody":
 "honnbunn"}]&automatic=false&transport=xmlhttp
```

図 7: 構造が特殊なパケット

4 フィルタリング手法の提案と設計

調査結果から, Web メールの使用を検出・制限するための解析・フィルタリングを行う手法を提案する. 本手法は, 以下の 2 つの部分から構成される.

(i) メール作成時に表示されるページの解析

(ii) メール送信時に送信されるパケットのペイロード部の解析

(i) は Web メールサイト内のメール作成時のページに 3-(1) で述べた特徴が現れているかを判別する. 検出対象となるメール特有のキーワードにはあらかじめ重みづけを行っておき, その値の合計が閾値を超えた時のみ規制する.

(ii) はメール送信時にクライアントから送信されたパケットのペイロード部に 3-(3) で述べた特徴が現れているかを判別する. (i) と同じく, キーワードには重み付けを行い, 値の合計が閾値を超えた時のみ規制する.

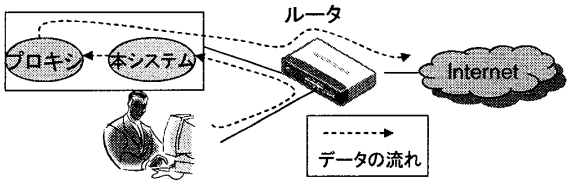

図 8: パケットの流れ

なお, 本システムは図 8 のようにプロキシ・クライアント間で流れているパケットを取得し, フィルタリングを行う.

5 評価実験

クライアントの Web ブラウザにプロキシとして本システムを設定し, 43 種類の Web メールサイト上でメールの作成・送信を行い, 検出・制限率の調査を行った. 今回の実験で利用したキーワードは表 1 であり, 実験結果は表 2 のようになった.

表 1: キーワード

	キーワード
手法(i)	(ii) + 宛先、件名、本文、添付
手法(ii)	from、to、cc、bcc、subject、body

表 2: 検出・制限率

	検出・制限率
手法(i)	81.4%
手法(ii)	88.4%
(i)+(ii)	97.7%

6 考察

(1) 本システムは (i) で規制することができない 3-(2) の特徴を持った Web メールを (ii) により規制することができる. また,(i) は (ii) と異なり, タグの属性以外のキーワードも検出するため, 3-(3) のキーワードの出現頻度が低いものも規制することができる. このように本システムではそれぞれのフィルタリング方式で規制できないものを残りの方式により規制することができるため, 評価実験の結果で示したような高い精度で Web メールを規制することが可能である.

(2) フィルタリング方式 (i) と (ii) を組み合わせても規制することができなかった 1 種類の Web メールサイトは 3-(2) と 3-(4) の特徴を持ったものであった. 図 7 のように現段階では規制することができない Web メールも存在しているが, このようなものが出現した際は個別に対処していく必要がある.

(3) 本システムは図 8 のような構成になっているため, クライアント側で本システムを中継するように設定する必要がある. しかし, 企業などのクライアント PC が多数存在する場合は, 設定を行う管理者の負担が大きくなると予想されるため, クライアントに設定を要求しない透過型プロキシのようなシステムが必要になってくると考えられる. また, 本システムは複数のクライアント・サーバ間の情報を処理しなければならないため, パケットごとの処理量をできるだけ少なくする必要がある.

7 今後の課題

現在, 本システムを実際に運用し, 一般的な Web サイト閲覧時の誤検出の発生率を調査中である. 今後は既存の方式を利用する場合と本システムを利用する場合の処理量の比較実験を行うと共に, さらに多数の Web メールサイトについての調査を行う予定である. また, 現状では https で暗号化された Web メールサイトを規制することができないため,https にも適用できるシステムの開発も検討している.

参考文献

[1] SECOM 情報漏洩対策サイト
http://www.secomtrust.net/infomeasure/rouei/column1.html

[2] S.Pukkawanna, V.Visoottiviseth and P.Pongpaibool: Classification of web-based email traffic in Thailand, *Proc. International Symposium on Communications and Information Technologies*, pp.440 - 445 (2006).