M_016

Web検索サービスを用いたウイルスメール収集手法

A Collection Method of Virus Mail using Web Search Engines

新井 貴之† 笹森 健司‡
Takayuki Arai Takeshi Sasamori

1. まえがき

インターネットの普及に伴ない, ウイルスメールやボットネットなどの不正プログラムの蔓延が問題になっている. これらの対策を検討するためには, まず検体の入手が課題になる. ウイルス検体を収集する手段としては, ハニーポットやネットワークモニタリングが知られているが, これらの手法では収集効率が設置するネットワーク環境に依存するため, 十分な収集が行なえない恐れもある. そこで本研究では, Web検索サービスを利用して, ウイルス検体が添付されたメールテキストを収集する手法について提案する.

2. 設計

コンピュータウイルスの9割以上がメール経由で感染していることが知られている[1]. ウイルスメールはメーリングリストなどによって感染を拡大する. 本研究では, インターネット上で公開されているメーリングリストのメールアーカイブに着目し, 検索エンジンを用いてウイルスメールを検索収集する手法を提案する. メーリングリストアーカイブには様々な形式があるが, 本研究では検体として扱いやすい mbox 形式のアーカイブファイルを収集対象とした. 収集にあたっては, 検索語や取得ファイルの選別を行ない重複検索や重複取得を排除する構造とした.

3. ウイルスメール収集手順

本研究では次の手順で自動的にウイルスメールの検索と収集を行なう手法を提案する.

(1) 初期キーワードの入手

ウイルスメールの検索に際しては, 特徴となるキーワードが必要である. ウイルスに関しては様々な情報があるが, 入手のしやすさや処理の簡便さなどから添付ファイル名を使用するものとした.

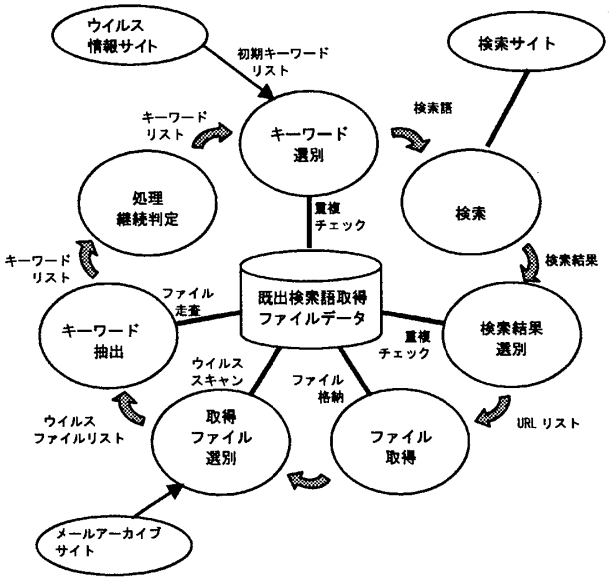
(2) 検索語の作成

入手したキーワードに検索条件を追加して検索語を作成する. mbox 形式で, 添付付きの, HTML 文書ではないメールを検索するため, 次の検索条件を使用した.

- 本文に "From " が含まれていること
- 本文に "Received: " が含まれていること
- 本文に "base64" が含まれていること
- 拡張子が ".mail" または ".mbox" であること

(3) 検索の実行

Web 検索サイトにアクセスし, 作成した検索語を用いて検索を実行して検索結果を得る. 検索に際してはWeb検索サイトの利用条件に従い, 濫用にならないよう検索頻度や検索回数に配慮する必要がある. 本システムでは, 検索実行前に既検索キーワードのチェックを行ない, 同一キーワードで重複検索を行なわない構造にする.



(4) 検索結果のスクリーニング

検索結果のうち, つぎの条件を満たすものを取得対象として選出する.

- 検索結果の見出しの先頭が "From " である
- 検索結果のダイジェストに検索キーワードが含れている

(5) ファイル取得の実行

作成した URL リストを用いて当該サイトからファイルを取得し, ハードディスクに格納する. 取得前に既取得ファイルのチェックを行ない, 同一 URL の重複取得を行なわない構造とした.

(6) 取得ファイルの選別

取得したファイルに対してウイルススキャンを実行し, ウイルスを含むファイルを選出してウイルスファイルリストを作成する. ウイルスが検出されなかったファイルについては, 添付ファイルの有無をチェックし分類する.

(7) キーワードの抽出

ウイルススキャンの結果ログから, ウイルスメールの添付ファイル名を抽出する. このキーワードを既出キーワードリストと照合し, 新出のものを新たなキーワードとする.

(8) キーワードの選別

† 横河電機株式会社, YOKOGAWA Electric Corporation

‡ 株式会社クルウィット, clwit, Inc.

抽出したキーワードをチェックし，検索語として不適なものを取り除く．具体的には，単語長が極端に短い(3 文字以下)ものや極端に長い(30 文字以上)もの，数字だけで構成されたもの，6 桁以上数字を含むもの(乱数生成された可能性がある)，ドメイン名を含むものなどを取り除く．

(9) 処理の継続判定

キーワードの選別の結果，新規のキーワードがある場合には，このキーワードを使用して上記の手順により再検索と再取得を実行する．この手順を新規キーワードが検出されなくなるまで繰り返すことにより，自動的に収集を行なう．

4. ウイルスメール収集システムの試作

提案手法を用いて，実際にウイルスメールを自動収集するシステムを試作した．

今回の試作では，Web 検索サービスに Google[2]を使用し，GoogleWebAPI[3]を用いて検索プログラムを実装した．収集ファイル選別のためのウイルス検知には，antivir[5]を使用した．検索に用いる初期キーワードは，ソフォス社の 2006 年 5 月のトップ 10 ウイルス[4]に挙げられているウイルスのうち，添付ファイル名記述のある 4 種類のウイルスの情報から取得した．取得結果を表 1 に示す．また，検索の繰り返し過程で得られた検索語数の推移を図 2 に，URL 数の推移を図 3 に示す．

表 1: ウイルスメール取得実験結果

	初期キーワード	取得処理繰り返し回数	キーワード総数	取得アーカイブ	ウイルスアーカイブ	ウイルスメール数	ウイルス種類数
NetSky.P	12	7	356	193	123	1507	70
NetSky.D	22	5	323	156	115	1446	62
Nyxem.D	23	8	349	173	119	1517	66
Mytob.M	8	5	323	154	114	1445	62
総計	63	--	406	193	127	1554	70

検索条件：「ファイル拡張子が mbox または mail」

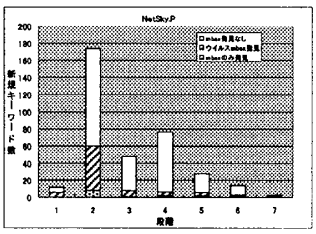


図 2：キーワード数の推移

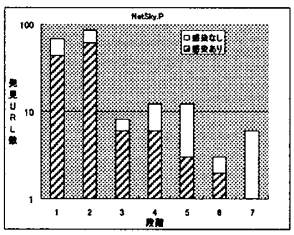


図 3：発見 URL 数の推移

表 1 の「総計」は，4 回の実験から重複を取り除いた値である．提案システムではキャッシュ機能により，既出キーワードは検索されず，既出 URL は取得されないため，「総計」項目のキーワード総数が実検索回数，取得アーカイブ数が実取得ファイル数になる．図 2 より，繰り返し過程で発見される単語のごく一部がウイルス mbox の URL を発見するためのキーワードになっていることがわかる．

図 3 より，発見された URL の過半数がウイルス mbox であった．なお，この実験で取得されたファイルの総量は約 70MB であり，所要時間は約 2 時間であった．処理時間のほとんどはファイル取得時間である．

この結果から異なる初期キーワードを用いて収集を行なっていても，取得結果の 8 割程度が重複していることがわかる．これは，取得処理の繰り返しによって，この検索条件下で取得可能なファイルの大半が取得されているためだと思われる．より多くのウイルスを取得するためには，初期キーワードを変更するよりも，検索条件を変更することが有効と考えられる．

検索条件のうち拡張子が".mbox"または".mail"を，".htm"または".html"ファイル以外に緩和して収集を行なったところ表 2 の結果を得た．

表 2: 検索条件を緩和した取得結果

	初期キーワード	取得処理繰り返し回数	キーワード総数	取得アーカイブ	ウイルスアーカイブ	ウイルスメール数	ウイルス種類数
NetSky.P	12	7	1189	329	329	8656	92
Nyxem.D	23	8	1202	330	330	8657	92
Mytob.M	10	8	1193	792	330	8657	92
NetSky.D	6	8	794	332	114	8664	96

検索条件：「拡張子が html または htm 以外」

検索条件を緩和すると，より多くのキーワードが発見されるため，処理の継続にはより多くの検索が必要となる．しかし，本システムでは，キーワードやファイル取得の重複取得を回避する機能があるため，徐々に検索条件を緩和することができれば，Web 検索サービスに大きな負荷をかけることなくウイルス収集を継続できるものと思われる．

まとめ

本稿では，ウイルス検体収集のための手段として Web 検索サービスを利用したウイルス収集手法を提案し，実装実験を通して，2 時間で 1500 通，70 種類程度のウイルスメールが収集できることを確認した．この手法を用いれば，ネットワーク環境に依存することなく，検体の収集を行なうことができる．今後は，検索条件の段階的な緩和方法や学習によるスクリーニングルールの変更，また，定期的な収集によるモニタリングへの応用についても検討する．

謝辞

本研究は，情報通信研究機構(NICT)から「広域モニタリングシステムに関する基盤技術の研究開発」として受託し，実施中である．ここに記して謝辞を表する．

参考文献

[1] 情報処理推進機構，コンピュータウイルスの届出状況 [2006 年 6 月分]について，http://www.ipa.go.jp/security/txt/2006/documents/virusfull0607.pdf

[2] Google, Inc.: Google, http://www.google.com/

[3] Goolge, Inc.: GoogleWebAPI, http://www.google.com/apis/index.html

[4] ソフォス:トップ 10 ウイルス, http://www.sophos.co.jp/security/top10/

[5] AVIRA: antivir, http:/www.free-av.com/