迷惑メール対策の現状と課題

2000MT034 泉澤 大資
指導教員　後藤 邦夫

1　はじめに

最近、PC や携帯電話に知らないアドレスからメールが来ることが増えてきた。内容は、アダルトサイトやチェーンメイルまで様々である。一日に数件程度ならよいが、何十件も来ると本来の有用なメールと区別しなければならず不便である。この迷惑メールの正体は「スパムメール」と言われる。現在のスパムの特徴として日に日に数が増えてきており、その種類も豊富になってきている。さらに英語のメールだけでなく日本語のメールも出始めている。本研究では、現在の対策から今後これらのスパムメールに有効となる対策について検討する。

2　スパムメールの現状

スパムの現状について述べる。

2.1　定義

『自分が望んでいない情報を一方的に送りつけてくるメール』のことをスパムメールという。

2.2　種類・目的

スパムは 2 種類に分類される。

1. UCE(Unsolicited Commercial Email) :受信者に望まれない商業宣伝メール
2. UBE(Unsolicited Bulk Email) :受信者に頼まれていない大量配信メール

スパムの目的を以下に示す。

1. アダルトホームページ宣伝
2. ねずみ講、マルチ商法
3. サイドビジネス勧誘
4. 商品の勧誘
5. 出会い系ホームページ勧誘

3　対策方法

スパムと非スパムを選別する「フィルタ」を用いた対策方法を検討する。

3.1　タイミング

メールの内容はユーザー毎に異なるため、その個人にあったユーザー独自のフィルタ作成が望ましい。そのため、MUA が受信した後に、フィルタリングする。

3.2　方法・手順

方法･手順を以下に示す。A Plan for Spam[1]、Better Bayesian Filtering[2] を使用した。

作成方法

メールの解析する対象を「ヘッダ、本文」とし、「バイナリの添付ファイル」は対象外とする。この対象情報から、ベイズの定理を用いてフィルタを作成する。確率が 0 でない事象 A と B に対して、条件付き確率の定理より、次式が成り立つ。

$$P\{A \mid B\} = \frac{P\{A\}P\{B \mid A\}}{P\{B\}} \tag{1}$$

このベイズの定理より、$B = (B \cap A) \cup (B \cap)$ と (B) と $(B \cap \bar{A})$ は互いに素な事象であるから次式を得る。

$$P\{A \mid B\} = \frac{P\{A\}P\{B \mid A\}}{P\{A\}P\{B \mid A\} + P\{\bar{A}\}P\{B \mid \bar{A}\}} \tag{2}$$

計算の手順

1. 準備
   大量のスパムメールと非スパムメールを用意し、それぞれに出現する単語の回数とメールを読み込んだ回数を数える。

   ヘッダ内の単語は属性をつけて表す。例えば “Subject” に含まれる単語であれば”Subject* 単語” とする。“o” を”O”、”a” を”@” 等の類語は、別の単語として数える。

2. 単語のスパム確率 P を求める。
   スパムメールに含まれる単語の割合と非スパムメールに含まれる単語の割合から判別したいメールの単語のスパム確率を計算する。

   - 読み込んだ (学習) スパム数 $(snum)$
   - 読み込んだ (学習) 非スパム数 $(cnum)$
   - 単語がスパムに現れた回数 (s)
   - 単語が非スパムに現れた回数 (c)

   誤検出を避けるために値を設定する。
   - 非スパム中の単語の回数を 2 倍にする。
   - 全体で 5 回以上出現していない単語は計算から外す。
   - 一度も現れていない単語のスパム確率を 0.4 とする。

- 一方の集合にのみ現れる単語の確率はそれぞれ 0.001 と 0.999 とする。

ベイズの定理 (2) 式を用いてスパム確率 P を求める。

$$P = \begin{cases} \frac{min(1.0, \frac{s}{snum})}{min(1.0, \frac{2c}{cnum}) + min(1.0, \frac{s}{snum})} & (2c + s > 5) \\ 0.4 & (others) \end{cases} \quad (3)$$

3. 結合確率を求める。
   メール中の 0.5 から最も離れている単語 (もっとも特徴的な単語)n 個のスパム確率を $(P_1$〜$P_n)$ とし、以下のように結合確率を求める。

   確率が 0 でなく互いに独立しており、かつ確率に対称性がある事象 $P_1, P_2, \cdots, n$ に対して次式が成り立つ。

$$\frac{P_1 * P_2 * \cdots * P_n}{P_1 * P_2 * \cdots * P_n + (1 - P_1) * (1 - P_2) * \cdots * (1 - P_n)} \quad (4)$$

   結合確率が 0.9 以上ならば、そのメールをスパムと判定する。

3.3 既存フィルタの手法と比較

- コンテンツ (ルールベースフィルタ)
  あるルールに従ってメールに点数をつけて判定する。あらかじめ点数が決まっているため、計算が高速である。しかし、新たな種類のスパムに対応できない。ベイズ理論のフィルタはデータに応じて自己修正できる。

- ホワイトリスト・ブラックリスト
  受信を許可、拒否するホストのリストを用いて判定する。他のフィルタと併用することで、計算を軽減することができる。

4 実装

日本語のスパムメールに対応するために日本語解析ツール「kakasi」を用いる。「kakasi」とは、漢字かなまじり文をひらがな文やローマ字文に変換し、単語毎に分割するツールである。

4.1 処理の流れ

学習 (データベースに保存)

1. 大量のスパムメールと非スパムメールを用意する。(日本語メールに対して「kakasi」を用いて日本語解析する。)
2. データベースを用意し、学習したメール数と単語の出現回数を保存する。
3. 計算式 (3) からスパム確率 P を求め、値をデータベースに保存する。

判定 (スパムと非スパムを判定)

1. 判別したいメールを単語毎に読み込む。
2. それぞれの単語のスパム確率 P をデータベースから参照し、特徴的な単語 n 個を選び出す。
3. 計算式 (4) から n 個の単語の確率を結合して判別する。
4. 判定結果に基づき、データベースを更新する。
5. スパムメールと非スパムメールを振り分ける。

5 評価

計算した単語のスパム確率 P の例を以下に示す。

words	スパム確率P		
研究室	0.001	※	0.999
検出	0.001	販売	0.999
時間	0.162627052384	こちらまで	0.999
方法	0.071013845466	希望	0.890410958904
次回	0.132610508757	ください	0.871794871794

図 1　日本語の単語のスパム確率

words	スパム確率P		
perl	0.001	madam	0.999
scripting	0.001	promotion	0.999
describe	0.001	cgi	0.9734398
example	0.033600237	enter	0.9075001
I'm	0.055427782	quality	0.8921298

図 2　英単語のスパム確率

6 おわりに

今後の改善策は、ベイズのフィルタは学習に応じて自己修正可能であり、より確実な確率を引き出せるということから、まず多くのメールを学習することである。また、計算中に設定した値の修正である。さらに、複数のフィルタを用いることによって、計算を軽減させることができる。このようにして、これから増えてくるであろう多種多様なスパムメールに対応できるのではないかと考える。

参考文献

[1] Paul Graham, A Plan for Spam, http://www.paulgraham.com/spam.html/, 2002.

[2] Paul Graham, Better Bayesian Filterring, http://www.paulgraham.com/better.html/, 2003.