5E-06

類似度を利用した迷惑メールフィルタリング

Spam mail filtering using similarity

渡邊　隆志†　佐藤　直†
Takashi　Watanabe　Naoshi Sato

1. まえがき

スパムメールを調査したところ，件名と本文が類似している，本文の文字数が少ない，本文中に URL が含まれている，といった特徴があることがわかった．そこで，件名と本文の類似度，本文の文字数，URL の有無という 4 つの特徴量を機械学習しスパム／非スパムを分類する手法を提案する．実際に受信した電子メールを対象に実験を行って提案の有効性を確認した．

2. 研究の背景，目的，関連研究

汎用的に使用されているメールクライアントソフトは，ベイジアンフィルタを搭載し，本文の内容（コンテンツ）を機械学習してスパムメールをフィルタリングする[1]．フィルタリングされたスパムメールは専用のフォルダに分類されることが多い．しかし，このコンテンツベースのフィルタリングには，単語の改変などにより容易にフィルタリングを通過できる，本文が短いスパムメールには対応できない，といった問題がある．実際，著者らの経験では，スパムメールを学習させたにもかかわらず，それらに類似した多くのスパムメールが非スパムメールとともに受信箱に残ってしまうという現象が見られる．　そこで，本研究は，スパムメールに関する本文の内容以外の特徴にも着目し，スパムメールの判定基準にすることでスパムメールを削減することを目的とする．

なお，関連研究として以下のような報告がある．すなわち，受信者の存在，送信者のドメイン，メールサイズ，URL の有無，同じ IP アドレスからの送信頻度，昼夜別の受信時刻といったスパムメールの行動の特徴からファジー決定木でスパム判定をする手法[2]，スパム／非スパムメール中で使用される単語の辞書をそれぞれ作成し，Jaro-Winkler 距離を使ってメール本文と辞書にある単語の距離を測定し同距離のマップを作成して判定する方法[3]，などがある．

以下では，従来の手法も参考に，類似度などの特徴抽出によるスパムメールフィルタリング法を検討する．

3. 電子メールの収集

著者の一人が使用している汎用 PC に搭載されているメールクライアントを利用して，約 10 ヶ月にわたって 21,128 件のメールを収集した．そのうち 5,000 件を非スパムメール，16,128 件をスパムメールと主観的に判定した．スパムメールのうち 15,594 件がメールクライアントに備わっているフィルタリング機能でフィルタリングされたスパムメール（以降，フィルタリングされたスパムメールと呼ぶ），534 件がフィルタリング機能を通過したスパムメール（以降，通過スパムメールと呼ぶ）である．

4. 通過スパムメールの特徴

スパムメールを目視したところ，件名と本文が類似しているものが多いことがわかった．また，本文中に URL が含まれていることや，本文の文字数が少ない，といった特徴があることも分かった．そこで，これらの特徴をスパム判定に用いることとする．件名や本文といった文字列の類似性を調べるアルゴリズムは，大きく文字ベースと単語ベースに分けられる[4]．単語ベースは形態素解析が必要になること，また，複数の区切り方がある場合や，未知語などが含まれている場合は解析精度が悪くなるため，ここでは文字ベースとする．文字ベースのアルゴリズムとしてよく利用される Jaro-Winkler 距離を用いる．

5. 特徴量

収集された電子メールについて前述の 4 つの特徴量を調べた．以下代表的な結果を示す．

（1）件名の類似度

Jaro-Winkler 距離を用いて件名の類似度を測った結果を図 1～図 3 に示す．なお，あるメールの類似度は他のメー

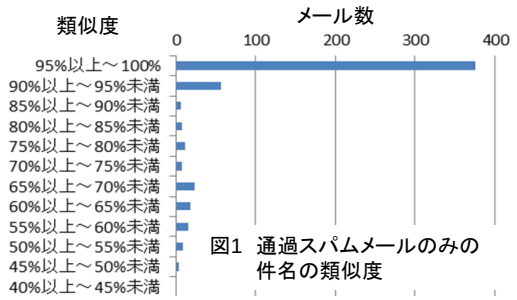

図1　通過スパムメールのみの件名の類似度

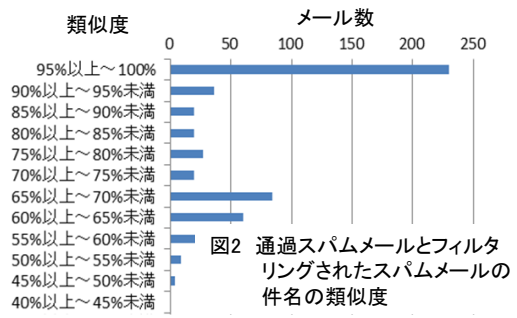

図2　通過スパムメールとフィルタリングされたスパムメールの件名の類似度

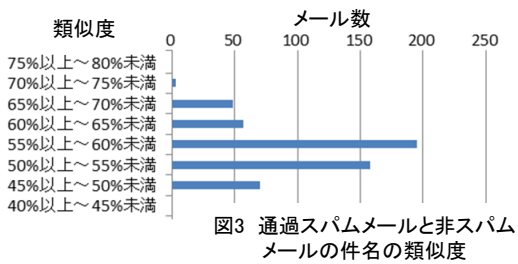

図3　通過スパムメールと非スパムメールの件名の類似度

† 情報セキュリティ大学院大学

ルとの間で計った Jaro-Winkler 距離の最大値である（以下同様）．図 1 は通過スパムメールのみのもので，類似度 95%以上のメールが大多数を占めていることがわかる．図 2 は通過スパムメールとフィルタリングされたスパムメールを合わせたもので，図 1 との比較から，フィルタリングされたスパムメールの類似度は 70%前後のものが多いことがわかる．図 3 は通過スパムメールと非スパムメールを合わせたもので，非スパムメールは通過スパムメールの 10 倍ほどあるため，同図から，非スパムメールの類似度は 55%前後のものが多いことが読み取れる．

（2）本文の類似度

同様に本文の類似度を測った結果を図 4～図 6 に示す．図 4 は通過スパムメールのみのもので，類似度 95%以上のメールが殆どであることがわかる．図 5 は通過スパムメールとフィルタリングされたスパムメールを合わせたもので，図 4 との比較から，フィルタリングされたスパムメールの類似度は 60%前後のものが多いことがわかる．図 6 は通過スパムメールと非スパムメールを合わせたもので，（1）と同じ理由で，非スパムメールの類似度は 50%前後のものが多いことがわかる．

（3）本文の文字数

本文の文字数を数えたところ，スパムメールは通過スパムメール，フィルタリングされたスパムメールともに最大 2000 字程度で 300 字未満のものが殆どを占めていることがわかった．一方，非スパムメールは最大 200000 字程度まで分布し，10000 字以内のものが多いことがわかった．

（4）URL の有無

本文中の URL の有無を調べたところ，スパムメールは通過スパムメール，フィルタリングされたスパムメールともに URL を含む傾向が強いのに対し，非スパムメールは URL を含まないものが多数であることがわかった．

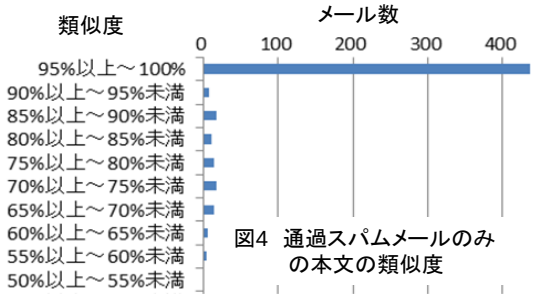

図4 通過スパムメールのみの本文の類似度

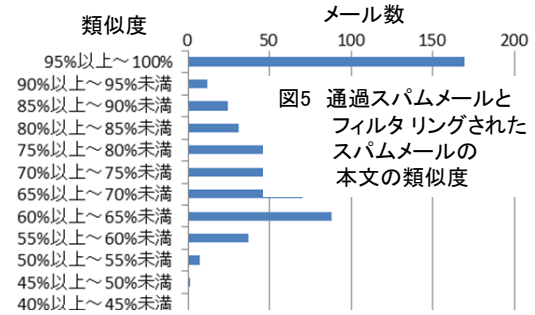

図5 通過スパムメールとフィルタリングされたスパムメールの本文の類似度

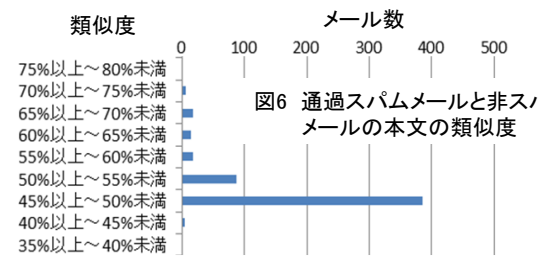

図6 通過スパムメールと非スパムメールの本文の類似度

6. 機械学習によるスパムメールの判別実験

以上から，スパムメールと非スパムメールの間には，4 つの特徴量について違いのあることがわかった．そこで，この 4 つの特徴量を用いた機械学習で判別する．この機械学習には，2 種類の識別性能に優れているサポートベクターマシン SVM を用いる．具体的にはツール LIBSVM[5]を用いる．**3** 章のように収集した全てのメールを対象に，コンテンツフィルタリングを経ずに，スパムメール／非スパムメールの判別実験を行った．スパムメールと非スパムメールをランダムに二分し，訓練データとテストデータとするスパムメール判別実験を実施した．結果を表 1 に示す．同表より，4 つの特徴量を全て用いた場合，非スパムメールの検出率（正解率）は 99.8%，スパムメールの検出率は 99.7%，となり，本提案が従来のコンテンツフィルタリング機能を代替できる見通しを得た．

7. むすび

本稿では，件名と本文の類似度，本文の文字数，本文中の URL の有無という 4 つの特徴量を使用して機械学習でメールを分類する手法を提案した．実験の結果，提案手法が既存のコンテンツフィルタリング機能を置換できる可能性があることを示した．

文　献

[1]田端利宏: SPAM メールフィルタリング : ベイジアンフィルタの解説，情報の科学と技術，Vol. 56, No. 10, pp. 464-468, 2006

[2]W.Meizhen, L.Zhitang and Z.Sheng: Fuzzy Decision Tree Based Inference Technology for Spam Behavior Recognition, ISPA, pp. 463-468, 2009.

[3]R.Ariaeinejad and A.Sadeghian: Spam detection system: A new approach based on interval type-2 fuzzy sets, Electrical and Computer Engineering, pp. 379-384, 2011.

[4]W.Gomaa and A.Fahmy: A Survey of Text Similarity Approaches, International Journal of Computer Applications, Vol. 68, No. 13, pp. 13-18, 2013.

[5]LIBSVM -- A Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/
（2015 年 1 月現在）

表1 機械学習によるスパムメール判別実験結果

ケース	SVMパラメータ（○は適用した特徴量） 件名の類似度	本文の類似度	本文の文字数	URLの有無	全スパム	実験 正解率（%）	総合正解率（%）
1	○				非スパム	96.5	96.7
					スパム	96.8	
2		○			非スパム	97.1	98.4
					スパム	98.8	
3	○	○	○	○	非スパム	99.8	99.7
					スパム	99.7	