フィルタ通過スパムメールの特徴と低減法の検討

渡邊 隆志† 佐藤 直†

†情報セキュリティ大学院大学
221-0835 神奈川県横浜市神奈川区鶴屋町 2-14-1

{mgs121501, sato}@iisec.ac.jp

あらまし スパムフィルタリング機能が搭載されているメールクライアントにスパムメールを学習させたにもかかわらず，スパムメールが非スパムメールとして受信フォルダに入ってしまうことがある．本検討は，このフィルタ通過スパムメールの特徴に着目し削減することを目的とする．フィルタ通過スパムメールには，件名と本文が類似している，本文の文字数が少ない，また，本文中にURLが含まれることが多い，といった特徴が見られる そこで，これらの特徴量を機械学習して分類する手法を提案する．実験の結果，提案手法，特に類似度に着目した判別法が有効であることを確認した．

Features and decrease method on unfiltered spam mails

Takashi Watanabe† Naoshi Sato†

† Institute of Information Security.
2-14-1, Tsuruya-cho, Kanagawa-ku, Yokohama-shi, Kanagawa-ken 221-0835, JAPAN
{mgs121501, sato}@iisec.ac.jp

Abstract Although mail users apply spam filter function based on Bayes' theorem, similar spam mails often pass the filter. In order to reduce the unfiltered spam mails, we surveyed the unfiltered spam mails and found they have the features as follows. That is, they have similar subject and body, the number of words in their body is small, and they include URL in the body, in comparison with normal mails. We propose a new method of identifying the unfiltered spam mails and decrease them by using Support Vector Machine. Through an experiment, we confirmed the effectiveness of the proposed method, especially usefulness of similarity.

1．はじめに

スパムメールの特徴に着目し，非スパムメールとして誤判定されたスパムメール（以下通過スパムメールと呼ぶ）を削減することを提案する．通過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通過スパムメールとして非スパムメールとともに受信フォルダに残ってしまうという現象が見られる．そこで，本検討は，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通過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．

しかし，本検討のように，通過スパムメールの低減に着目したフィルタリング手法の検討は見受けられない．そこで，以下では，従来の手法も参考に，特徴抽出による通過スパムメール低減法を検討する．

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，534 件が通過スパムメールである．

4. 通過スパムメールの特徴

通過スパムメールを目視したところ，フィルタリングされたスパムメールとの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．

なお，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．本文では，代表的な距離尺度として，文字列の類似度評価によく利用される 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．

5.1 類似度

最初に，本検討における件名や本文の類似度を以下のように定義する．すなわち，件名や本文の全体を一つの文字列とみなし，Jaro-Winkler 距離を用いて，比較する件名や本文との類似度を算出する．算出された類似度の最大値を同件名や本文の類似度とする．類似度は 0 から 1 の範囲（0～100%）で与えられる．なお，Jaro-Winkler 距離の測定には Java ライブラリ Apache Lucene[5]を使用する．

（1）件名の類似度

上述のように定義した類似度を用いて件名の類似度を測った．結果を図 1～図 3 に示す．

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

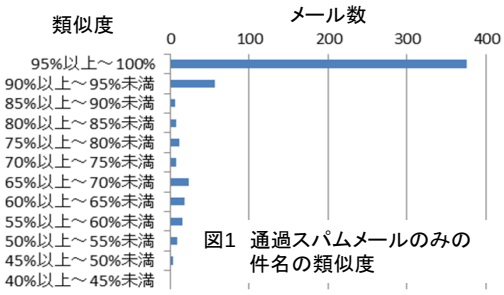

図1　通過スパムメールのみの件名の類似度

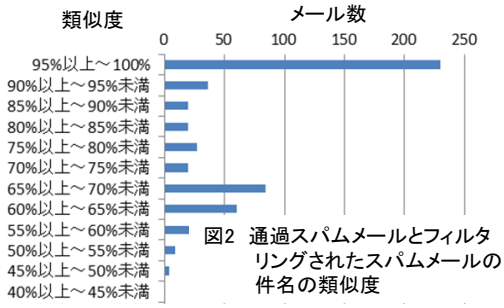

図2　通過スパムメールとフィルタリングされたスパムメールの件名の類似度

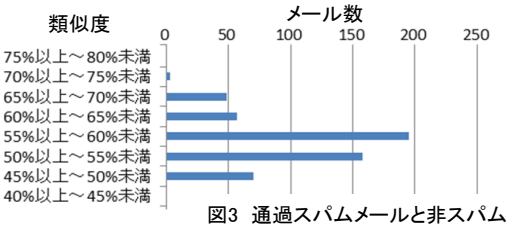

図3　通過スパムメールと非スパムメールの件名の類似度

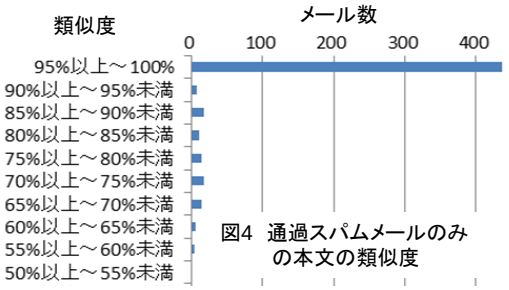

図4　通過スパムメールのみの本文の類似度

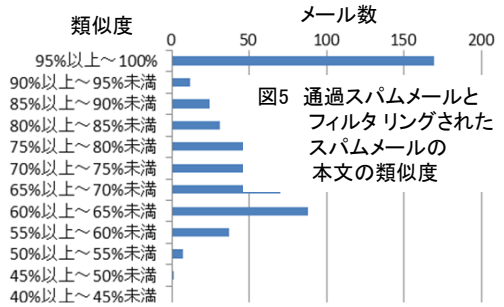

図5　通過スパムメールとフィルタリングされたスパムメールの本文の類似度

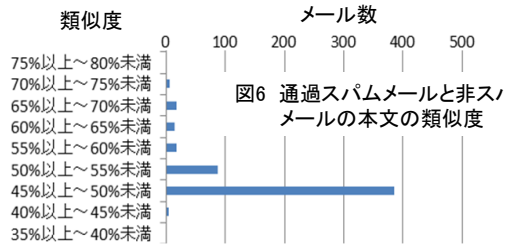

図6　通過スパムメールと非スパムメールの本文の類似度

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.

5.2 本文の文字数

通過スパムメール, 非スパムメールの本文の文字数の累積分布をそれぞれ図7, 図8に示す (図中の実線. 通過スパム全体. 非スパム全体). なお, 両図の横軸の補助目盛は{2, 3, 5. 7}×10^nを表す.

図 7 より, 通過スパムメールの本文は高々 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る.

図 8 より, 非スパムメールの本文は 200000 字程度まで広く分布し, 平均 10000 字程度であることがわかった.

5.3 URL の有無

本文中の URL の有無を調べたところ, 通過スパムメールの 96%, 非スパムメールの 37%が URL を含んでいた. 本文中に URL を通過スパムメール, 非スパムメールの本文の文字数の累積分布をそれぞれ図7, 図8に示す (図中の破線. URL を含む通過スパム. URL を含む非スパム).

図 7 より, 通過スパムメールの殆どが URL を含むことが確認できる.

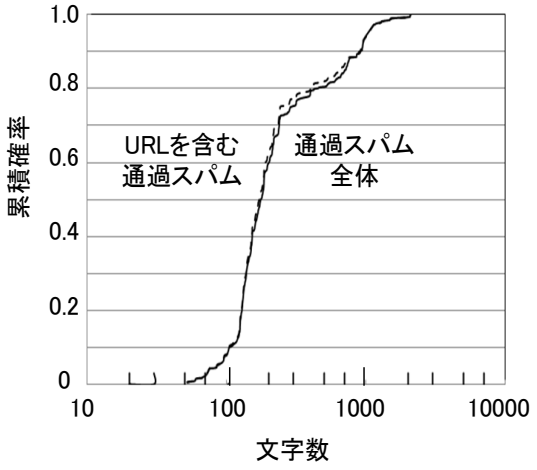

図7 通過スパムメール本文の文字数

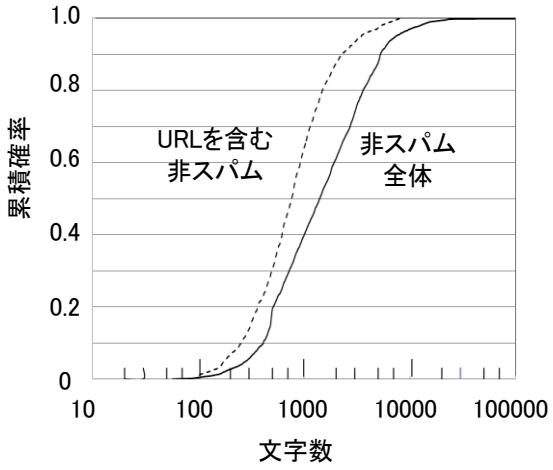

図8 非スパムメール本文の文字数

図 8 から，URL を含む非スパムメールは，非スパムメール全体の中では，平均的に文字数が少ないことも分かる．

以上の結果から，4 つの特徴量のうち，通過スパムメールと非スパムメールの間で大きく異なる特徴量は件名と本文の類似度であることがわかった．本文の文字数については，通過スパムメールの文字数が比較的に少ないということが分かった．ただし，非スパムメールの文字数が広く分布していることから，通過スパムメール・非スパムメール判別に用いる特徴量としては類似度ほど強くはないといえる．URL の有無については，通過スパムメールの殆どが URL を含むことが分かった．ただし，非スパムの 37%が URL を含むことから，通過スパムメール・非スパムメール判別に用いる特徴量としては，本文の文字数と同様，強くはないといえる．

6．通過スパムメールの判別実験

以上の検討から，通過スパムメールと非スパムメールの間には，件名と本文の類似度について，大きな違いのあることがわかった．

そこで，件名と本文の類似度を主たる特徴量として，同一の受信フォルダに収容されている両者を機械学習し判別する．また，本文の文字数，URL の有無も特徴量として機械学習し判別に用いることとする．

多くの機械学習の手法が提案されているが，本検討では，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 を用いることとする．具体的にはツール LIBSVM[6]を用いる．以下，前述のように収集した通過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して，通過スパムメール判別実験に用いる．

通過スパムメールの判別実験結果を表1の実験 1 として示す．実験では，特徴量の組み合わせを 12 通りに変えて（SVM パラメータ．ケース 1～ケース 12）実施した．

同表の正解率とは，非スパム／通過スパムそれぞれを正しく判別（検出）する確率であり，総合正解率とは，メール数で両正解率を荷重した値である．

同表から，機械学習に用いる特徴量を増やすと正解率が向上することが読み取れる．4 特徴量を全て用いた場合（ケース 12），通過スパムメールの正解率は 95.1%，非スパムメールの正解率は 99.9%であった．これは，現在の，コンテンツフィルタリングを経て受信メールフォルダに保存されるメールに本提案を適用することによって，非スパムメールとして誤判定され保存されるスパムメールの約 95%が低減可能であることを意味する．一方，本提案は非スパムメールの 0.1%をスパムメールとして誤判定するリスクがあることも分かる．

表1　実験結果

ケース	SVMパラメータ				非スパム／通過スパム（実験1）／全スパム（実験2）	実験1		実験2	
	件名の類似度	本文の類似度	本文の文字数	URLの有無		正解率（%）	総合正解率（%）	正解率（%）	総合正解率（%）
1	○				非スパム	99.6	96.2	96.5	96.7
					通過スパム／全スパム	64.8		96.8	
2		○			非スパム	99.9	98.7	97.1	98.4
					通過スパム／全スパム	87.3		98.8	
3	○	○			非スパム	99.9	98.7	99.5	99.4
					通過スパム／全スパム	88.0		99.4	
4	○		○		非スパム	99.7	96.9	95.9	97.3
					通過スパム／全スパム	70.8		97.8	
5	○			○	非スパム	99.8	96.5	95.4	97.1
					通過スパム／全スパム	65.5		97.6	
6		○	○		非スパム	99.8	98.6	98.8	99.2
					通過スパム／全スパム	86.5		99.4	
7		○		○	非スパム	99.9	98.8	97.5	98.6
					通過スパム／全スパム	88.4		98.9	
8	○	○	○		非スパム	99.9	98.8	99.8	99.6
					通過スパム／全スパム	89.1		99.5	
9	○	○		○	非スパム	99.9	99.1	99.3	99.5
					通過スパム／全スパム	91.4		99.5	
10	○		○	○	非スパム	99.9	98.2	96.4	97.9
					通過スパム／全スパム	84.6		98.4	
11		○	○	○	非スパム	99.7	99.3	99.2	99.3
					通過スパム／全スパム	95.1		99.3	
12	○	○	○	○	非スパム	99.9	99.4	99.8	99.7
					通過スパム／全スパム	95.1		99.7	

なお，本検討は，通過スパムメールの低減を図る目的で実施した．上記判別実験が比較的良好な結果を示したことから，参考のため，収集した全てのメールを対象に，すなわち，コンテンツフィルタリングを経なかった場合を想定して，同様にスパムメール／非スパムメールの判別実験を行った．その結果，表1の実験2として示す．実験 2 の結果から，4 特徴量を全て用いた場合（ケース 12），スパムメール（全スパム）の正解率は99.6%，非スパムメールの検出率は99.8%となり，本提案が従来のコンテンツベースのフィルタリングを代替できるという見通しを得た．

また，実験1と実験2を比較すると，実験2のスパムメールの正解率が向上している．すなわち，表1において，実験1の通過スパムよりも実験2の全スパムの正解率が大きい．これは，メール数の違いから，実験 1 に比べ実験 2 では，SVM の内部パラメータが非スパムメールの判定よりもスパムメールの判定にチューニングされた結果であろう．

なお，判別に要する計算時間はどのケースも概ね数分程度であった．

7. むすび

本検討は，通過スパムメールを削減することを目的に，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，既存のコンテンツフィルタリング機能に提案手法を追加することによって，通過スパムメールを大幅に削減できることを確認した．また，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を示した．なお，機械学習期間と判別特性の関係の分析は今後の課題である．

参考文献

[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]Apache Lucene: http://lucene.apache.org/ (2014/8)

[6]LIBSVM -- A Library for Support Vector Machines: http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (2014/8)