複数内容を考慮したスパムメールフィルタリング

佐々木 稔　新納 浩幸
茨城大学工学部

1　はじめに

近年，スパム (Spam) メールがネットワーク社会において重要な問題となっている．このスパムメールは，送信するコストが安いこともあり，欲しいかどうかにかかわらず不特定多数の人に向けて同じメールが一方的に送られる．そのため，スパムメールが大量に送られると，スパムメールと必要なメールの選別をする作業に時間を取られる．また，スパムメールがメールサーバに不具合を起こさせることもある．スパムメールの問題を解決するために，これまで様々な方法によりスパムメールを規制する取り組みがなされている．

このスパムメール問題を解決するために，クライアント側でスパムメールかどうかを判定し，フィルタリングを行う研究が盛んに行われている．しかし，現在ではスパムメールの中にもいくつかの種類が存在するため，フィルタリング技術を用いたとしても，必要なメールであると判定するスパムメールも存在する．このような現状を改善するため，我々のグループは複数のスパムメールの内容を考慮して，スパムメールの判定を行う研究を行っている．この研究は，スパム，非スパムすべてのメールを，一般的なクラスタリングアルゴリズムである k-means アルゴリズムを用いて自動的にいくつかのクラスタに分類することで，様々な内容を持つスパムメールを個別の内容としてとらえることを目的としている．

これまでの研究において，本研究で提案した手法は SVM(Support Vector Machine) を利用したフィルタリングと同等の性能が得られることが分かっている．しかし，スパムメールのフィルタリングで最も重要なことは，必要なメールを誤ってスパムメールと判断してはいけないことである．これは，誤判別によって大量のスパムメールから必要なメールを探す手間が生じることを防ぐためである．これまでの研究では，必要なメールの誤判別が少なからず存在しているため，実用に耐えうる研究結果が出ているとはいえない．クラスタリングを行う際，クラスタにスパムメールと必要なメールが混在していることで，スパムか必要なメールかのラベル付けに原因があるのではないかと考えられる．これまでクラスタに含まれるスパムメールの割合を 7 割以上と設定していたが，本稿ではこの閾値を様々な値に設定し実験を行い，それによる精度の変化を調査する．

2　スパムメールの現状

スパムメールは，不特定多数の人に向けて主に営利目的やいたずら目的で大量に配信されるメールで，スパムの他にもジャンクメール (Junk Mail)，UBE(Unsolicited Bulk Email) や UCE(Unsolicited Commercial Email)，日本では迷惑メールなどと呼ばれている．現在，スパムメールにもいくつかの種類が存在する．例えば，薬物の購入やお金儲けなどを勧誘する広告や偽りのウワサを広めることを目的とした都市伝説と呼ばれるメールやチェーンメールなどがある．以前は，ポルノ画像などのサイトを紹介するメールやコンピュータウイルスが添付されたメールが大部分を占めていたが，ここ数年にかけて薬物や違法コピーされたソフトウェアの広告などに移り変わっている．また，HTML メールに見えないほどの小さな画像を張り付けておくことにより，送り主がアクティブなユーザであることを確認する Web Bug と呼ばれるスパムメールも存在する．

最近では，広告やチェーンメールなどのいたずらが目的のメールの他にフィッシング (Phishing) と呼ばれる悪質な詐欺目的のメールも存在する．これはクレジットカード会社などのメールや Web サイトを巧妙に偽造し，パスワードやクレジットカード番号，個人情報などを盗み取るものである．本物との区別が付かないためにこれまでにもフィッシングに

よる被害が深刻化し，日本でも平成 16 年 11 月に被害届が出されている．そのため，今後この問題への何らかの対策が急務となっている．

このように，スパムメールの種類は年々増加しており，送付する目的も広告やいたずらから詐欺行為へと悪質化している．また，その手口も巧妙化しているため送付者の特定や不正中継などでメールを送信したサーバの特定も非常に難しくなっている．

3　スパムメール問題への取組み

このようなスパムメールの問題を解決するために，これまで様々な方法によりスパムメールを規制する取り組みがなされている．この取り組みは大きく 2 つに分けることができる．ひとつは，政治的に法律などを定めてスパムメールを規制することである．アメリカ合衆国では，いわゆる「スパム規制法案」が 2003 年に可決され，スパムメールの送信行為を厳格に規定している．日本では，携帯電話を中心にいわゆる「迷惑メール」を規制する「迷惑メール防止法案」が 2002 年に可決されている．しかし，法的に規制をしても完全にスパムメールの問題が解決された訳ではなく，法律での規制に存在する抜け穴を利用してスパムメールを送るために現在でもスパムメールの問題は根強く残っている．

もうひとつは，技術的にスパムメールをフィルタリングすることである．これは，送られてきたメールのアドレス，ヘッダ，メールの内容を解析して，スパムメールと判断されたものにはそのメールを見なくても済むようにカラー表示などのマークを付けることでユーザの手助けを行う．技術的な方法でスパムメールをフィルタリングするには，サーバレベルとクライアントレベルの 2 か所で対策を行うことができる．サーバレベルでのスパムメール対策としては，SMTP または MTA においてスパムメールを送ろうとするスパマーが送信時にスパムメールを送ろうとするのをブロックするものがある．スパムメールを受信する時は，サーバ側でスパマーが使っているであろう IP アドレスのリストを参照して，リストに存在している MTA からはスパムメールであると判断する [5]．このリストを作成する作業は時間の経過とともに増加するために，手作業でのリスト作成は手間のかかる作業となる．そのため，現在では SMTP relay サーバリストを作成，公開している団体も存在している．

クライアント側でスパムメールのフィルタリングを行う研究は近年盛んに行われ，現在では Microsoft の Outlook や Mozilla の Thunderbird など数多くのシステムにこのフィルタリング機能が実装されている．これまでに提案されたフィルタリング技術には，C4.5 [8]，Ripper [4] や Support Vector Machine(SVM) [6, 7] などの機械学習手法や Naive Bayes を用いた確率モデル [1, 9] が存在する．この中で，Naive Bayes を用いたスパムメールのフィルタリング手法は簡単な学習で高い精度でスパムメールを判定するので，上記のメールクライアントの他，多くのフィルタリングツールにおいて採用されている．ただ，これまでの研究では，スパムか非スパムかを判別するために 2 つの事後確率や頻度分布を学習し，判別するものが主流である．

4　スパムメール判定手法

現在のスパムメール判定手法を用いる場合，スパムメールと非スパムメールのそれぞれに対応する 2 つの単語頻度統計を計算することによってスパムメールを判定するモデルを作成している．そのため，同じような内容を持つスパムメールに関しては精度良く判定をすることができるが，あまり受け取らないものは頻度の少なさから誤ってスパムメールではないと判定する可能性がある．このことから，多様化するスパムメールの内容をひとつの頻度統計で表現するのは，受け取ったメールの判断が非常に難しくなるのではないかと考えられる．また，多様化するスパムメールへの対応を可能とするため，これまで学習した判定モデルに加えて，新しく受け取ったメールに対して動的な判定モデルの更新を行うことが望ましい．

これらの問題点を解決するスパムメール判定モデルについて説明する．本研究では，複数のスパムメールの内容を考慮して判定を行う手法を提案する．この手法は，まずすべてのメールを一般的なクラスタリングアルゴリズムである球面 k-means アルゴリズムを用いて指定した数のクラスタに分類する．次に，得られたクラスタの中にスパムメールが存在する割合を計算し，クラスタがスパムであるか非スパムであるかをラベリングする．これまでに行った実験では，クラスタ内に 7 割以上のスパムメール

が存在する場合，そのクラスタはスパムであるとラベリングしていた．実験を行った結果，スパムメールを高い判定率でフィルタリングできたが，必要なメールをスパムメールであると誤判別するメールが存在している．このような誤判別を少なくするためには，スパムメールの判定率は減少するが，閾値を上げることで 100%まで上げることができる．本実験では，閾値を上げて判別実験を行うことでスパムメールの判定率がどの程度変化するのかを調査する．

クラスタのラベルが決定されると，受け取ったメールをスパムメールかどうか判定するため，クラスタをひとつのベクトルに変換する．受け取ったメールをベクトルに変換することにより，メールとクラスタの類似度計算を容易に行うことが可能となる．これまでは Naive Bayes 手法や SVM を用いた手法などではスパムメールのトピックをひとつの単語統計で表現していたが，スパムメールが持つ複数のトピックをモデル化することが可能となる．

スパム，非スパムを表す概念ベクトルが得られると，受け取ったメールがどのクラスタに最も近いかを類似度の計算を行うことでランク付けを行う．受け取ったメールを表すベクトルと概念ベクトルとの類似度には，ベクトルの余弦 (cosine) を用いてベクトル間の類似度を計算する [2] ．すべての概念ベクトルについて余弦計算を行い，計算結果が最も大きいクラスタを求め，そのクラスタのラベルをメールの判定結果として返す．

5 実験

本節では，前節において述べたスパムメール判定手法の評価実験を行い，本手法の有効性を検証する．本実験で用いた実験データには Ling-spam を用いた．Ling-spam のデータは全体で 2893 件のメールが存在し，この中で 481 件はスパムメール，2412 件は non-spam メールに分かれている．また Ling-spam には，元のメール文書の他に，停止語 (stop word) を削除したデータ，見出し語変換 (lemmatizer) を行ったデータ，停止語辞書と見出し語変換の両方を行ったデータの合計 4 種類のデータが用意されている．この 4 種類のテストセットに対して，全体の 90% (non-spam: 2170 件，spam: 432 件) を学習データとし，残りの 10%(non-spam: 242 件，spam: 49 件) をテストデータとしてスパムメール判定実験を行った．

メールの本文を計算機で扱える表現とするために，文書検索などで使われるベクトル空間モデルを使う．メールをベクトル化する際，ベクトルの要素には出現する単語の重みが使われる．単語の重みを計算する方法はいくつか提案されているが，本研究では TF·IDF 法と TF のみ使った方法を用いた [3]．TF·IDF 法では，j 番目のメールに出現する i 番目のタームの重みを以下のように表す．

$$w_{ij} = f_{ij} \cdot \log \frac{n}{f_i}$$

ここで，n はメールの全体数，f_{ij} は j 番目のメールに出現する i 番目の単語の頻度，f_i はメール全体における i 番目の単語が出現する文書数を表す．これにより，すべてのメールをメールベクトルとして表現する．得られたベクトルに対してクラスタリングを行い，自動的に 50 個のクラスタに分割し，閾値に応じてスパムクラスタと非スパムクラスタとラベリングする．

5.1 実験結果・考察

さまざまな閾値を設定して実験を行った結果を表 1 と表 2 に示す．この表において，Data はデータの種類を表し，bare は生のメールデータ，stop は停止語を削除したデータ，lemm は見出し語処理を行ったデータ，lemm_stop は停止語の削除と見出し語処理をともに行ったデータである．また，Threshold はスパムとラベリングするための閾値，Spam Precision はテスト用スパムメールの正解率，Non-Spam Precision はテスト用非スパムメールの正解率を表す．

表 1，2 より，TF·IDF，TF ともに高い精度でスパムメールを判定していることが分かる．スパムメールの判定のみを比較すると，lemm_stop を除いて TF を用いる方が TF·IDF より高い閾値精度でも精度が高い．クラスタリングを行った際にスパムクラスタにはスパムメールが大部分を占めていることから，モデル化と判定が共に精度の高いレベルとなっている．

非スパムメールの判定精度に注目すると，テストデータ 242 件中，誤った判別をしたものが 0〜2 個であった．このデータに関しては 100%の正解率が望まれるため，誤判別のないように閾値を上げると，TF·IDF では高かったスパムメールの判定精度が大

表 1: 閾値を変化させたときの実験結果 (TFIDF 使用)

Data	Threshold	Spam Precision	Non-Spam Precision
bare	0.8	95.92%	99.59%
bare	0.85	87.76%	100.00%
lemm	0.7	95.92%	98.76%
lemm	0.8	71.43%	100.00%
stop	0.8	91.84%	99.59%
stop	0.85	89.80%	100.00%
lemm_stop	0.7	97.96%	98.76%
lemm_stop	0.75	83.67%	100.00%

表 2: 閾値を変化させたときの実験結果 (TF 使用)

Data	Threshold	Spam Precision	Non-Spam Precision
bare	0.8	95.92%	99.59%
bare	0.85	95.92%	100.00%
lemm	0.7	97.96%	99.17%
lemm	0.8	93.88%	100.00%
stop	0.8	95.92%	99.59%
stop	0.85	95.92%	99.59%
lemm_stop	0.8	89.80%	99.17%
lemm_stop	0.85	75.51%	99.17%

幅に下がる結果となった．しかし，TF のみを用いると精度の大幅な減少はなく，約 94% 以上の高いスパムメール判定精度が得られた．TF·IDF を用いた場合，テスト用スパムメールを判定すると少数のスパムメールが含まれる非スパムクラスタに最も近いと判定され，誤判別を引き起こしてしまう傾向があった．Ling-spam の中には，スパムと区別の付きにくいメールも多く含まれているため，スパム，非スパムメールが混在したクラスタも存在している．このような場合でも，非スパムメールを正しく判別できるように改良を行うことは今後の課題となっている．

6 おわりに

本稿では，複数内容のスパムメールを考慮したスパムメール判定研究において，ラベル付けの際の閾値を変化させたときの判定精度の変化を検証した．その結果，非スパムメールを正しく判別できるように閾値を変化しても，高い精度でスパムメールを判別できることが分かった．今後は，スパムメールの判別精度がさらに良くなるように改良を行っていく予定である．

参考文献

[1] I. Androutsopoulos, J. Koutsias, K. Chandrinos, G. Paliouras, and C. Spyropoulos. An evaluation of naive bayesian anti-spam filtering. In *Proceedings of the Workshop on Machine Learning in the New Information Age: 11th European Conference on Machine Learning (ECML 2000)*, pages 9–17, 2000.

[2] M. W. Berry and M. Browne. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. SIAM Book Series: Software, Environments, and Tools, 1999.

[3] M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using linear algebra for intelligent information retrieval. In *SIAM Review*, volume 37, pages 573–595, 1995.

[4] W. W. Cohen. Learning rules that classify e-mail. In *Proceedings of the 1996 AAAI Spring Symposium on Machine Learning in Information Access*.

[5] O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *SIGMOD Rec.*, 30(4):55–64, 2001.

[6] H. Druker. Support vector machines for spam categorization. In *Proceedings of the IEEE Transaction on Neural Networks*, volume 10, pages 1048–1054, 1999.

[7] A. Kolcz and J. Alspector. Svm-based filtering of e-mail spam with content-specific misclassification costs. In *Proceedings of the TextDM '01 Workshop on Text Mining, IEEE International Conference on Data Mining*.

[8] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[9] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk e-mail. In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*.