社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ＦＩ－ 76　(11)
2004－ＮＬ－163　(11)
2004／9／16

文書分類を用いたスパムメール判定手法

佐々木 稔†　　新納 浩幸‡

†茨城大学 工学部 情報工学科 ‡茨城大学 工学部 システム工学科

†‡〒316-8511　茨城県日立市中成沢町 4-12-1

† sasaki@cis.ibaraki.ac.jp ‡ shinnou@dse.ibaraki.ac.jp

概要

近年，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．そこで，本研究で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手法を提案する．この手法は，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．その結果，我々の提案手法はスパムメールを約 90% 以上，非スパムメールを 96% 以上の高い精度で判別することができた．そのため，スパムメールが持つ広範囲なトピックを抽出することやスパムメールの細かい特徴をとらえることが可能となった．また，SVM と bogofilter を用いて比較評価を行ったところ，提案手法は SVM と比較して少々見劣りはするものの同等の判定精度を持ち，bogofilter と比較すると非常に有効な手法であることが分かった．

Automatic Spam Detection Method Using Text Clustering

Minoru Sasaki†　　Hiroyuki Shinnou‡

†Department of Computer and Information Sciences, Faculty of Engineering, Ibaraki University

‡Department of Systems Engineering, Faculty of Engineering, Ibaraki University

4-12-1, Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan

† sasaki@cis.ibaraki.ac.jp ‡ shinnou@dse.ibaraki.ac.jp

Abstract

In recent years, spam email is a widespread problem on the Internet. When a large number of spam messages are recieved, it is necessary to take along time to identify spam or non-spam email. To solve the spam problems, there have been several attempts to detect and filter the spam email on the client-side. However, almost approaches learn and find the distribution of the feature set in only the spam and the non-spam messages. In this paper, we propose a new spam detection technique using document clustering algorithm based on vector space model. Our method computes disjoint clusters automatically using a spherical k-maens algorithm for all spam/non-spam emails and obtains centroid vectors of the clusters. We describe the our spam detection system and show the result of our experiments using the Ling-Spam test collection.

1 はじめに

近年，スパム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，送られた人はスパムメールと必要なメールの選別をする作業に多大な時間を取られる．また，このスパムメールが原因となり、メールサーバに不具合を起こさせることもある．

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 [6]．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，盛んに行われている．これまでの研究では，C4.5 [13]，Ripper [5] や Support Vector Machine(SVM) [8, 12] などの機械学習手法や Naive Bayes を用いた確率モデル [2, 14] が提案されている．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，最近では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．

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．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．最近では，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．そのため，このようなフィルタリング技術を用いたとしても誤って判定するスパムメールも数多く存在している．

そこで，本研究で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手法を提案する．この手法は，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．これにより，スパムメールが持つ広範囲なトピックを抽出することが可能となり，高い精度でスパムメールを判別することが期待できる．また，スパム，非スパムメールの特徴をベクトルとして表現しているので，細かい特徴をとらえることが可能となる．

第 2 節では，これまでに提案されたスパムメール

判定手法の問題点を指摘し，それを改良した新しい判定手法の提案を行う．第 3 節では，提案手法による判別実験を行い，その結果と他の手法との比較，評価を行い，第 4 節でまとめを行う．

2　スパムメール判定手法

本節では，これまでに提案されたスパムメール判定手法の問題点を指摘し，それを改良した新しい判定手法の提案を行う．

2.1　スパムメール判定手法の問題点

現在スパマーが送るスパムメールの種類も多様化し，スパムメールフィルタをすり抜けるスパムメールが次々と送られている．これまでは，スパムメールのほとんどが薬物の広告であったが，最近ではチェーンメール，都市伝説と呼ばれる誤った噂話，宝くじに当選したという偽りの当選報告など，単にスパムメールと言ってもトピックの範囲は非常に広くなっている．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．これまでに学習を行った判定モデルを用いて，非スパムメールをスパムメールと誤って判定した場合，そのメールがスパムメールだと判定するように学習をする必要がある．しかし，追加学習をする際，これまでに学習したモデルを捨てて新しくモデルを構築するのは，モデルを作成する時間が必要となる．全体的なメールの数が少ない時は再構築でもモデル作成が可能であるが，メールの数が大規模になるほどモデル作成に必要な時間が多くなる．また，モデルの作成時に学習データとして使ったメールを保存しているとは限らないため，追加学習をしても以前のモデルを復元することができず，適切に更新できない可能性がある．そのため，動的な文書分類手法や関連性フィードバックなど，これまでに作成したモデルを容易に更新できる判定モデルを採用すべきだと考えられる．

2.2　提案手法

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．この球面 k-means アルゴリズムについては，2.4 節において詳しく述べる．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．このとき，クラスタ中のスパムメールの割合は具体的には定めていないが，本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．クラスタの代表ベクトルを求めるために，本研究では球面 k-means アルゴリズムで得られる概念ベクトルをクラスタの代表ベクトルとする．概念ベクトルについては，2.3 節で述べる．これにより，これまで 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広範囲なトピックを表現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 [3] ．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．

2.3 概念ベクトル

ベクトルの集合をベクトル空間にプロットしたとき，同質のベクトルが多く存在する場合を除いて，いくつかのグループに分かれる．このようなグループはクラスタと呼ばれ，類似した内容をもつベクトルの集合が形成される．概念ベクトルはクラスタに属するベクトルの重心を求めることにより得られ，そのクラスタの内容を表す代表ベクトルである．

概念ベクトルを求める例として，正規化された N 個のベクトル $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ を，異なる s $(s < N)$ 個のクラスタ $\pi_1, \pi_2, \cdots, \pi_s$ にクラスタリングすることを考える．このとき，ひとつのクラスタ π_j に含まれるベクトル x_i の平均である重心 $\mathbf{m}_j$ は以下のように表される．

$$\mathbf{m}_j = \frac{1}{n_j} \sum_{\mathbf{x}_i \in \pi_j} \mathbf{x}_i \tag{1}$$

ここで n_j はクラスタ π_j に含まれるベクトルの数を表す．ベクトルの重心は単位長にはなっていないので，そのベクトルの長さで割ることにより概念ベクトル $\mathbf{c}_j$ を得る．

$$\mathbf{c}_j = \frac{\mathbf{m}_j}{\|\mathbf{m}_j\|} \tag{2}$$

2.4 球面 k 平均アルゴリズム

k 平均アルゴリズムでは，目的関数は一般的に概念ベクトルとクラスタに属するベクトルとの距離の和

$$\sum_{\mathbf{x}_i \in \pi_j} \|\mathbf{m}_j - \mathbf{x}_i\| \tag{3}$$

を最小にするような概念ベクトルを求める，最小二乗法が用いられる．球面 k 平均アルゴリズムでは，このような最小化問題ではなく，ミクロ経済学の分野における，生産計画の最適化問題で扱われている目的関数を用いている [11]．これは，各クラスタ $\pi_j (1 \leq j \leq s)$ の密度を

$$\sum_{\mathbf{x}_i \in \pi_j} \mathbf{x}_i^T \mathbf{c}_j \tag{4}$$

とし，クラスタの結合密度の和を目的関数としている．

$$D = \sum_{j=1}^{s} \sum_{\mathbf{x}_i \in \pi_j} \mathbf{x}_i^T \mathbf{c}_j \tag{5}$$

この目的関数 D を最大にするように，ベクトルの集合を反復法によりクラスタリングする．文書ベクトル $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ を s 個のクラスタ $\pi_1^\star, \pi_2^\star, \cdots, \pi_s^\star$ に分割するためのアルゴリズムを以下に示す．

1. すべての文書ベクトルを s 個のクラスタに任意に分割する．これらの部分集合を $\{\pi_j^{(0)}\}_{j=1}^s$ とし，これより求められた概念ベクトルの初期集合を $\{\mathbf{c}_j^{(0)}\}_{j=1}^s$ とする．また，t を繰り返しの回数とし，初期値は $t = 0$ である．

2. 各文書ベクトル $\mathbf{x}_i (1 \leq i \leq N)$ に対し，余弦が最も大きい，最も文書ベクトルに近い概念ベクトルを見つける．このとき，すべての概念ベクトルは正規化されているので，余弦は文書ベクトル $\mathbf{x}_i$ と 概念ベクトル $\mathbf{c}_j^{(t)}$ の内積を求めることと同値である．これにより，前回の繰り返しで求めた概念ベクトル $\{\mathbf{c}_j^{(t)}\}_{j=1}^s$ から，文書ベクトルが新たな部分集合 $\{\pi_j^{(t+1)}\}_{j=1}^s$ に分割される．

$$\begin{aligned} \pi_j^{(t+1)} &= \{\mathbf{x}_i : \mathbf{x}_i^T \mathbf{c}_j^{(t)} \geq \mathbf{x}_i^T \mathbf{c}_l^{(t)}\} \\ &\quad (1 \leq l \leq N,\ 1 \leq j \leq s) \end{aligned} \tag{6}$$

ここで，$\pi_j^{(t+1)}$ は概念ベクトル $\mathbf{c}_j^{(t)}$ に近いすべての文書ベクトルの集合とする．

3. 新たに導かれた概念ベクトルの長さを正規化する．

$$\mathbf{c}_j^{(t+1)} = \frac{\mathbf{m}_j^{(t+1)}}{||\mathbf{m}_j^{(t+1)}||}, \quad (1 \leq j \leq s) \tag{7}$$

ここで，$\mathbf{m}_j^{(t+1)}$ はクラスタ $\pi_j^{(t+1)}$ の文書ベクトルの重心を表す．

4. 目的関数 $D^{(t+1)}$ の値を求め．前回の繰り返しにおける目的関数の値 $D^{(t)}$ との差を計算する．このとき，

$$\|D^{(t)} - D^{(t+1)}\| \leq 1 \qquad (8)$$

を満たす場合，$\pi_j^\star = \pi_j^{(t+1)}$，$\mathbf{c}_j^\star = \mathbf{c}_j^{(t+1)}$ $(1 \leq j \leq s)$ とし，アルゴリズムを終了する．停止基準を超えていない場合は，t に 1 を加え，ステップ 2 に戻る．ここで，停止基準における目的関数の差は，文書数が約 4000 で，クラスタの数が 8 よりも大きい場合，収束した時の目的関数は 1000 を超えることがこれまでの研究で報告されている [7]．このため，繰り返しでの 1 以下の差は無視できるとし，便宜的に 1 という値を設定した．

3 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．

3.1 データ

本実験では，スパムメール判定に用いた実験データとして，テストコレクションのひとつである Ling-Spam [1] [2] を利用した．Ling-Spam のデータは全体で 2893 件のメールが存在し，その内，481 件のスパムメールが存在し，2412 件の 非スパム メールが存在する．また Ling-Spam には，元のメール文書の他に，メール文書からあらかじめ停止語を削除したデータ集合，メール文書の各単語に対して見出し語変換を行ったデータ集合，停止語の削除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．これらの 4 種類のテストコレクションに対して，それぞれ全体の 90% (非スパム: 2170 件，スパム: 432 件) を学習データとし，残りの 10%(非スパム: 242 件，スパム: 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．このとき，テストデータに対する判別精度の計算方法には，非スパムメールのテストデータが判別により正解した割合，スパムメールのテストデータが判別により正解した割合を判別精度とした．

[1] http://idl.ils.unc.edu/~efrom/data/lingspam/

3.2 実験方法

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 法を用いた [4]．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．

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

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，df_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した．

得られたメールベクトルに対し，本研究での中心的な部分となるスパムメール，非スパムメールにおける様々な内容を考慮するため，単語についての頻度統計を計算する．これを実現するために，すべてのメールを対象に球面 k 平均アルゴリズムを利用して，指定されたクラスタ数にクラスタリングを行う．このとき，得られた各クラスタに対して，そのクラスタがスパムメールであるか非スパムメールであるかをラベリングする必要がある．本実験では，便宜的に学習時に与えたスパムメールが，クラスタ内に存在するすべてのメール数の 70% 以上存在していれば，そのクラスタはスパムメールを表しているものとみなす．それ以外は非スパムメールを表すクラスタであるとする．これにより，クラスタの集合をスパムクラスタと非スパムクラスタに分類することが可能となり，受け取ったメールとの比較をするために，クラスタの重心ベクトルを計算し，ラベルに応じてスパムベクトル，非スパムベクトルとする．

スパムベクトルと非スパムベクトルが得られると，受け取った新しいメールがスパムであるか非スパムであるかをベクトル空間内で判断することができる．

Filter Config	Cluster	Spam Precision	Non-Spam Precision
bare	50	91.84%	99.17%
bare	100	89.80%	99.59%
lemm	50	95.92%	98.76%
lemm	100	95.92%	97.52%
stop	50	93.88%	99.17%
stop	100	95.92%	98.35%
lemm+stop	50	97.96%	98.76%
lemm+stop	100	100%	96.28%

表 1: 提案手法における実験結果

受け取ったメールはクラスタを計算した際のベクトル化と同様にして，ベクトル空間内でベクトルとして表現し，スパムベクトルと非スパムベクトルの内積をそれぞれ計算する．内積計算の結果，類似度が最も大きくなるベクトルがスパム，非スパムのうちどちらであるかによって，新しいメールのラベリングを行う．

3.3 実験結果・考察

この実験の結果を表 1 に示す．この表において，“Cluster” はクラスタリングを行う際のクラスタ数で，今回の実験では 50 と 100 の 2 つの場合について判定精度を求めた．また，“spam precision” と “non-spam precision” はそれぞれスパムメール，非スパムメールの正解率を表す．この表から，スパムメールは約 90% 以上，非スパムメールは 96% 以上の精度でスパムメールを判別し，高い判別精度を得ることができた．複数のスパム，非スパムクラスタを用いることでも十分高精度なフィルタリングが可能である．このことから，本研究におけるモデルが多様化する スパムメールにも十分対応できる柔軟性を持つことが分かった．しかし，球面 k 平均アルゴリズムを用いる場合，ひとつのクラスタには大多数のスパムメールの他にいくつかの 非スパムメールが存在することがある．そのため，それらのメールに存在する単語分布があるスパムメールの単語分布に悪影響を及ぼしていると考えられる．そのため，非スパムメールが 100% 必要なメールだと判定できるように，Rocchio の手法などのような関連性フィードバック手法を用いて個々のクラスタをチューニングすることが今後の課題となる．

次に，提案手法の精度を他の手法を用いた精度と比較することにより，提案手法の客観的な評価を行う．ここでは，機械学習手法で最も注目されている Support Vector Machine(SVM) [10] と現在メールクライアントなどで最も広く使われている Naive Bayes 手法 [9] を用いたフリーのスパムメール判定ツールである bogofilter [1] の 2 つを用いて判定を行った結果を示す．

まず初めに，SVM を用いたスパムメール判定実験について述べる．SVM に入力するメールベクトルは提案手法で作成したものと同じ TF·IDF 法を用いて作成した．また，SVM を利用するにあたり，本実験では SVMlight [10] を用いた．得られたメールベクトルに対してあらかじめ与えられたスパムと非スパムのラベルを付与し，SVM を使って学習を行った．このとき，SVM のパラメータに関しては特に何も指定せず，また SVMlight のオプションについても全く記述せず，標準の設定において実験を行った．

次に，bogofilter を用いたスパムメール判定実験について述べる．bogofilter は，メールを直接コマンド入力をしてデータベースに登録するため，1 通のメールを保存したファイル名とそのラベルを引数

Filter Config	SVM		bogofilter	
	Spam Precision	Non-Spam Precision	Spam Precision	Non-Spam Precision
bare	97.96%	100%	36.73%	100%
lemm	97.96%	100%	42.86%	100%
stop	97.96%	100%	36.73%	100%
lemm+stop	100%	100%	40.82%	100%

表 2: SVM, bogofilter における実験結果

として実行する．これをすべての学習データについてコマンド入力を行い，スパムと非スパムの頻度統計を計算する．bogofilter についても SVM と同様に様々なオプションが用意されているが，本実験ではオプションとして何もしていない，デフォルト設定のもとでメール内容の学習を行った．

この実験結果を表 2 に示す。非スパムメールのテストデータに対しては，SVM，bogofilter ともに誤って判定することなく，100% の正解率で正しく判定した．自動的なスパムメール判定ツールを作成する際，非スパムメールを誤判定することは必要なメールを見失う可能性があるために，避けなければいけない．これら 2 つの手法は，非スパムメールをひとつの内容としてとらえることで，多くの学習データを使ってモデルが作られているために高精度の認識が可能となっていると考えられる．提案手法においても，非スパムメールの誤判定が 0% となるように，非スパムメールをひとつの内容と考えてひとつのベクトルとして表現するなど，更なる精度向上を目指して工夫をする必要がある．

スパムメールの判定精度に関しては，SVM は 98% 以上の高い精度で判定ができるのに対し，bogofilter では 40% 前後の精度となった．SVM は様々な分野において非常に強力な学習能力を示し，その有効性はすでに実証されている．スパムメール判定に対しても，SVM が少ない学習データに対してスパムメールと非スパムメールとの境界を作っていることがわかる．bogofilter については，我々が問題点として指摘したスパムメール全体をひとつの内容として頻度統計を計算していることで，様々な内容をもつテストデータに対応できなかったのではないかと考えられる．これらの手法を提案手法と比較すると，提案手法は bogofilter と比較して非常に高い判定精度があり，SVM と比較して少々見劣りはするものの同等の判定精度を持っていることが分かった．しかし，提案手法ではモデルを修正することが容易なので，メールを受け取っても判定モデルの再構築をする必要がないが，SVM ではメールを受け取るたびに判定モデルを再構築する必要がある．最近では，SVM の処理を高速に行うことが可能であるので再構築のコストもそれほど高くはないが，あらかじめ存在する判定モデルを修正して新しいメールに対応するのもひとつの方法として利用可能ではないかと考えられる．そのため，適合性フィードバックや動的なクラスタリング手法を使うなど，判定モデルの効率的な修正方法についての工夫が今後の課題となる．

4 おわりに

本稿では，近年多様化するスパムメールの内容を考慮して，文書クラスタリング手法を用いたスパムメール判定手法を提案した．その結果，我々の提案手法はスパムメールを約 90% 以上，非スパムメールを 96% 以上の高い精度で判別することができた．そのため，スパムメールが持つ広範囲なトピックを抽出することやスパムメールの細かい特徴をとらえることが可能となった．また，SVM と bogofilter を用いて比較評価を行ったところ，提案手法は SVM と比較して少々見劣りはするものの同等の判定精度を持ち，bogofilter と比較すると非常に有効な手法

であることが分かった．現在，SVM の処理は高速化が進んでいるので判定モデルの再構築にかかるコストもそれほど高くないが，提案手法は判定モデルの再構築，現在ある判定モデルの修正とどちらの処理も効率的に行うことが可能であるため，スパムメール判定モデル構築の一手法として利用可能ではないかと考えられる．

今後の課題としては，提案手法ではすべてのメールをクラスタリングし，その後でラベリングを行っているので，ひとつのクラスタに大多数のスパムメールと少数の非スパムメールが存在する場合がある．そのため，Rocchio の手法などのような関連性フィードバック手法を用いて個々のクラスタをチューニングし，あらかじめ存在する判定モデルを修正して新しいメールに対応することが挙げられる．また，現段階では非スパムメールを判定した正解率が 100% に達していないので，どのデータに対しても 100% の精度を持つ判定モデルを構築する必要がある．それには，非スパムメール全体をひとつのベクトルとして表現して判定モデルを構築する，また，個々の誤判別事例からモデルのチューニング方法を探るなどして，高い精度を実現する方針としている．

参考文献

[1] *Bogofilter*. http://bogofilter.sourceforge.net/.

[2] 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.

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

[4] 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.

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

[6] 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.

[7] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. Technical report, IBM Almaden Research Center, 1999.

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

[9] P. Graham. *Better Bayesian Filtering*. http://www.paulgraham.com/better.html.

[10] T. Joachims. *Learning to Classify Text Using Support Vector Machines*. Dissertation, Kluwer, 2002.

[11] J. Kleinberg, C. Papadimitriou, and P. Raghavan. Segmentation problems: A micro-economic view of data mining. In *Proceedings of the 30th ACM Symposium on Theory of Computing*, pages 473–482, 1998.

[12] 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*.

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

[14] 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*.