博士論文

辞書にない語のスパムメール分類性能解析と応用手法の開発

(Characterization of Strange Words for Spam Mail Classification and Development of Application Methods)

令和5年10月

天満　誠也

山口大学大学院創成科学研究科

学位論文要旨

学位論文題目	辞書にない語のスパムメール分類性能解析と 応用手法の開発 (Characterization of Strange Words for Spam Mail Classification and Development of Application Methods)
指導教員	松野　浩嗣　教授
申請者名	天満　誠也 山口大学大学院創成科学研究科 自然科学系専攻

スパムメールを自動的に分類するため，これまで多くの機械学習に基づくメールフィルタリング手法が提案されているが，完全フィルタリングには至っていない．この原因の一つに，スパム送信者がフィルタすり抜けのために作成した，単語を記号，スペース及びHTMLタグ等の組み込みによって改変した文字列による影響がある．例えば，スパムメールには「price$ for be$t drug$!」,「priceC I A L I S」,「< font> se< / font> xu< font> al< / font >」等が含まれており，これらは，記号等の組み合わせの変更により，日々新しい文字列に置き換えられている．

機械学習に基づくフィルタリング手法では，過去に受信したメール群での単語の出現傾向を捉え，これらを分類対象のメールが含む単語に当てはめることでメールを分類する．上記のような改変された文字列の中には，学習用と分類用メール群の両方に出現する，すなわち未処理のまま利用することで，スパムメールの特徴として使えそうな語を含む一方で，分類用メール群にのみ出現する，すなわち機械学習ができておらず，利用のために特別な処理（記号等の削除や，類似語の探索等）が必要となる語も含んでいるが，既存手法ではこれらの区分をしておらず，同一の処理を施して扱っている．

そこで，上記のような改変された文字列を，学習用と分類用メール群の両方に出現する語と，分類用メール群にのみ出現する語に区分し，そ

れぞれ分類に利用する手法を開発することで，既存手法の分類性能を向上させ，完全フィルタリングにより近づけることを試みる．本研究では，上記のような改変された文字列を，多くのフィルタリング手法で用いている形態素解析システムによる扱いに合わせ,「辞書にない語」として扱う．この辞書にない語の典型例には，上記の他に，正規メールが含む新語，親しい間柄で用いる固有名詞，及び略語等がある．

本研究で得られた成果は以下の通りである．

(1) 辞書にない語と辞書に載っている語の分類性能を比較するため，既存手法を用い，辞書にない語，名詞，動詞及び形容詞のそれぞれを用いた分類実験を行った．その結果，辞書にない語の分類性能が最も高いことがわかった．すなわち，辞書にない語が分類性能に与える影響が大きいということであり，これの利用手法を開発することで，既存フィルタリング手法の分類性能のさらなる向上が期待できる．

(2) 辞書にない語の内訳について調べるため，学習用と分類用メール群の両方に出現する語と，分類用メール群にのみ出現する語の種類数を各々集計し，その結果を，同様の集計を行った名詞，動詞及び形容詞での結果と比較した．その結果，辞書にない語には，その他の品詞と比べて，学習用と分類用メール群の両方に出現する語のうち，正規またはスパムメールの一方にのみ出現する，すなわち分類に最も役に立つ出現パターンの語がかなり多いことがわかった．他方で，分類用メール群にのみ出現する，すなわち機械学習できない語もかなり多いことがわかった．これらを区分し，利用手法をそれぞれ開発することで，分類性能の向上が期待できる．

(3) 辞書にない語の利用について，(A) 学習用と分類用メール群の両方に出現する語の利用手法と，(B) 分類用メール群にのみ出現する語の利用手法をそれぞれ開発した．

(A) 学習用と分類用メール群の両方に出現する語の内訳について，正規またはスパムメールの一方にのみ出現する，すなわち分類性能が向上する出現パターンの語とそうでない語に分け，メールの件名や本文で用いられた数を各々で調べた．その結果，分類性能が向上する出現パターンの語は，そうでない語よりも多くのメールで用いられる傾向にあることがわかった．すなわち，用いられるメールの数のしきい値を定め，これを超える語を分類に用いることで，分類性能が向上する出現パターン

の語を分類に多く利用できるということである．これを実行する手法を開発し，しきい値を変えながら実験を行い，最適な値を探索した結果，7付近に設定することで，分類性能が向上することを確かめた．

(B) 分類用メール群にのみ出現する辞書にない語の種類数について，正規とスパムメールで比較した結果，正規よりもスパムメールの方が多い傾向にあることがわかった．この差を分類に利用するため，分類用メール群にのみ出現する辞書にない語に対し，スパム確率を一律に設定し，分類実験を行った．その結果，スパム確率を 0.7 に設定することで，分類精度が 98.2%から 98.9%向上した．

上記 (A)，(B) を併用することで，辞書にない語のうち，学習用と分類用メール群の両方に出現する語と，分類用メール群にのみ出現する語の両方を分類に利用でき，精度向上を得ることができる．

メールフィルタリングは改良が重ねられ，その性能は限界まで来ている．さらなる精度向上，すなわち完全フィルタリングに近づけるためには，新しい観点が必要であり，本論文は辞書なし語の利用という，その観点のひとつを与えるものである．

本論文は以下のように構成される．

第 1 章では，機械学習を用いたメールフィルタリング手法の背景について説明し，このようなフィルタをすり抜けるため，スパム送信者が辞書にない語を用いることについて述べたうえで，本論文の目的及び構成を述べる．

第 2 章では，これまでに提案されたフィルタリング手法を例に挙げ，関連研究について説明する．

第 3 章では，準備として，使用する電子メールのデータセット，及び単語の取り扱いと辞書なし語について説明したのちに，分類性能の評価に用いる尺度である ROC 曲線，データの集まり具合やばらつき具合を調べるための散布図，及び箱ひげ図について説明を行う．

第 4 章では，辞書なし語とそれ以外の単語で分類性能を比較した結果から，辞書なし語が分類性能に与える影響が大きいことを示す．さらに，辞書なし語の内訳について調べた結果から，学習用と分類用メール群の両方に出現する語と，分類用メール群にのみ出現する語が両方とも多いことを示し，これらを分けて扱うことで，分類性能が向上する可能性について述べる．次章以降でこれに取り組み，結果を報告する．

第５章では，先に述べた(A)，すなわち辞書なし語のうち，学習用と分類用メール群の両方に出現する語の利用手法を開発する．単語ごとに，メールの件名や本文で用いられた数を集計した結果から，その数が，分類性能を低下させる語ほど少ない傾向にあることを示す．この結果を基に，メールの件名や本文で用いられた数にしきい値を定め，それを超える語のみを分類に用いる手法を提案する．しきい値を変えながら実験をうことで最適な値を求め，これによる性能向上の効果について報告する．

第6章では，先に述べた(B)，すなわち辞書なし語のうち，分類用メール群にのみ出現する語の利用手法を開発する．この単語の種類数について，正規とスパムメールでそれぞれ調べて比較した結果から，スパムメールのほうが多い傾向にあることを示したうえで，この特徴が，スパムメール検出のためのバイアスとして利用できることを示す．本論文ではbsfilterと併用する実験について扱い，辞書なし語のうち，分類用メール群にのみ出現する語に対してスパム確率を一律に設定する手法を開発し，最適なスパム確率を探索した結果，スパム確率0.7で分類性能が大きく向上することを報告する．

第７章では，５章で開発した学習用と分類用の両方に出現する語の利用手法と，６章で開発した分類用メール群にのみ出現する語の利用手法を組み合わせた処理の流れについて説明し，今後の展望を述べ，本論文をまとめる．

Abstract

Title of Thesis	Characterization of Strange Words for Spam Mail Classification and Development of Application Methods)
Name of degree candidate	Seiya Temma
Degree and Year	Ph.D. in Science, 2023
Thesis directed by	Dr. Hiroshi Matsuno Professor Graduate School of Sciences and Technology for Innovation Yamaguchi University, Japan

Many mail filtering methods have been proposed, but they have not yet achieved perfect filtering. One of the reasons for this is the influence of modified words created by spammers to slip through the mail filtering, in which words are modified by insert symbols, spaces, HTML tags, etc. For example, " price$ for be$t drug$! ", " priceC I A L I S ", " <font> se</ font> xu< font> al</ font> " , etc. These are frequently replaced with new strings by changing the combination of symbols ,HTML tags etc.

Mail filtering is a technique that captures trends in words in training mails (mails received in the past) and applies these trends to words in test mails (newly received emails). Some of the above modified words appear in both training and test mails, i.e., words that could be used as features of spam mail by using them unprocessed, while others appear only in test mails, i.e., words that have not been learned and require special processing (e.g., removal of symbols, search for similar words, etc.) for their use. However, existing methods do not make these distinctions and treat them in the same way.

Therefore, in order to bring the filtering performance of the existing

methods closer to perfect filtering, we developed a method in which the above modified words are separated into words that appear in both training and test mails and words that appear only in test mails, and each of these words is used for mail filtering.

In this study, we treat the above modified words as "strange words". Typical examples of such strange words include, in addition to the above, new words included in ham mails, proper nouns used in close relationships, and abbreviations.

The results of this study are as follows

(1) In order to compare the filtering performance between strange words and other words, filtering experiments were conducted using existing methods with strange words, nouns, verbs, and adjectives. The results showed that the filtering performance of the strange words was the best. This means that strange words have a significant impact on the filtering performance, and we expect to improve the filtering performance of existing methods by developing a new method to utilize strange words.

(2) In order to examine the breakdown of strange words, we counted the number of words that appeared in both training and test mails, and the number of words that appeared only in test mails. The results were compared with those obtained for nouns, verbs and adjectives. We found that there are a significant number of strange words that appear in both training and test mails, but only in one of the groups, i.e., ham or spam mail. Words with this appearance pattern are most useful for mail filtering. On the other hand, we found that there are many strange words that appear only in test mails, i.e., words that cannot be learned. We expect to improve the filtering performance by separating these strange words and developing a new method to use each of them.

(3) For the use of strange words, we developed (A) a method for using words that appear in both training and test mails, and (B) a method for using words that appear only in test mails, respectively.

(A) To examine the breakdown of strange words that appear in both training and test mails, we divided them into two categories: words that appear only in ham and spam mails, i.e., words with patterns that im-

prove filtering performance, and words that do not, and examined their frequency of occurrence. The results showed that the words with appearance patterns that improve filtering performance tend to appear more frequently than those without such patterns. This means that by using words with a certain number of occurrences in filtering, it is possible to use more words that improve filtering performance. We developed a method to do this and conducted experiments with different threshold values to find the optimal value, and confirmed that setting the threshold around 7 improves filtering performance.

(B) We compared the number of strange words that appear only in the test mails between ham and spam mails, and found that the number tends to be higher in spam mail than in ham mail. In order to utilize this difference for filtering, we proposed a method to set a uniform spam probability for strange words that appear only in the test mails, and attempted to find the optimal spam probability. As a result, setting the spam probability to 0.7 improved the filtering accuracy from 98.2% to 98.9%.

By using (A) and (B) above together, both words that appear in both training and test mails and words that appear only in test mails can be used for mail filtering to increase accuracy.

Mail filtering has been improved and its performance has reached its limit. In order to further improve accuracy, i.e., to approach perfect filtering, a new perspective is needed, and this paper provides one such perspective: the use of strange words.

This paper is organized as follows.

In Chapter 1, we review the background of mail filtering methods, discuss how spammers use strange words to slip through such filters. The purpose and structure of this paper are then presented.

In Chapter 2, we will discuss related research on examples of filtering methods that have been proposed so far are given.

In Chapter 3, we describe the mail datasets, word handling, and strange words used in the this paper. This is followed by an explanation of the ROC curve, which is the measure used to evaluate the filtering perfor-

mance, and explanation of scatter plots and box-and-whisker plots.

In Chapter 4, we compare the filtering performance between strange words and other words, and show that strange words have a significant impact on the filtering performance. Furthermore, based on the results of a breakdown of the number of strange words, we discuss the possibility of improving filtering performance by separating words that appear in both training and test mails from those that appear only in the test mails. We will work on this in the next chapters and report the results.

In Chapter 5, we develop a method to use (A) above, i.e., strange words that appear in both training and test mails. From the results of counting the number of words used in the subject and body of each email, we show that the number tends to be smaller for words that degrade the filtering performance. Based on these results, we propose a method that sets a threshold for the number of words used in the subject and body of mails, and uses only those words that exceed the threshold for classification. Experiments are conducted to find the optimal value by varying the threshold, and the effect of this method on performance is reported.

In Chapter 6, we develop a method to use (B) above, i.e., strange words that appear only in the test mails. We compare the number of types of these words in ham and spam mails, and show that the number tends to be larger in spam mails, and that this feature can be used as a bias for detecting spam mails. In this paper, we deal with experiments using bsfilter and develop a method to set spam probabilities uniformly for strange words that appear only in the test mails. After searching for the optimal spam probability, we report that a spam probability of 0.7 greatly improves the filtering performance.

In Chapter 7, we describes the processing flow combining the methods developed in Chapter 5 and Chapter 6. The paper is then summarized, including future prospects.

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　本研究の背景

スパムメールは未だ減少しておらず，電子メール全体の約4割を占めており，その被害額も増加傾向にある [1]．この対策のために，これまでに多くのフィルタリング手法が開発されており，文献 [2] では，これらを non-AI based と AI based に分けたうえで，詳しく比較している．

non-AI based には，メール送信に用いたサーバーの認証や，ホワイトリストとブラックリストを用いた分類手法があるが，スパム送信者が行う，ヘッダ情報の改ざんや，スパムメールの内容を変化させる行為によってすり抜けられることがあり，これに対応するためには，分類のために必要な情報を，人手によってこまめに更新する労力がかかってしまう．

AI based では，過去に受信したメール群から，スパムメール分類に必要な特徴抽出を自動的に行うことができる．この機械学習を用いたフィルタリング手法は，現在広く研究されており，完全フィルタリングに近い分類が可能となっている．

スパムメールの分類のために最初に提案されたフィルタリング手法は，Paul Graham 提案の方式 [3] のベイジアンフィルタ [4] である．これは，過去に受信したメール群に出現する単語の出現頻度を，正規とスパムメールに分けて集計し，これらの比を単語のスパム確率として学習する．

メールを分類するとき，分類対象のメールが含む単語に対して，学習した単語のスパム確率を当てはめ，これらを掛け合わせることによってメールのスパム確率を求めている．

これを改良した手法が Gary Robinson 提案の方式 [16] である．この手法では，学習用メール群での出現頻度が少ない単語ほど，分類性能に与える影響を小さくするようなバイアスをかけている．これにより，単語の出現頻度が十分でない，いわゆるノイズのようなものの悪影響を少な

くすることができる.

さらにこれを改良した手法が Gary Robinson-Fisher 方式 [17] である. この手法では, メールのスパム確率を求めるために, 各々の単語について, カイ 2 乗分布を用いた検定を行うことによって分類性能を向上させており, bsfilter[18] や, Thunderbird[19] といったメールソフトなどで現在広く用いられている.

ベイジアンフィルタ以外にも, 決定木を用いた分類手法である機械学習システム BONSAI を用いた手法や, サポートベクトルマシン (SVM) を用いた分類手法など, 多くの手法がこれまでに提案されている.

機械学習システム BONSAI[9] は, 一次元の記号列データを対象にし, 正の例と負の例の二つの学習例から, それらを分類するための規則を決定木の形式で導き出す機能を持つ.

杉井ら [11] は, BONSAI を用いてメールを分類する手法を開発している. これは, メールヘッダと本文が含む各々の単語を, 正規とスパムメールでの出現数の偏り具合に応じた記号文字に変換し, メールデータを一次元の記号列データに変換することで BONSAI に適用し, 高性能な分類を可能としている.

SVM[7] は, 正と負の 2 つの学習例を高次元空間に付置し, それらを分離する超平面を求めることで, 線形識別を行う手法である.

平ら [30] は, これをテキスト分類に適用するため, 単語数を次元とするベクトルによって文章を表現する手法を提案している.

米倉ら [20] は, これを電子メール分類へ適用した効果について調べており, 高性能な分類ができることを確かめている.

上記の手法の多くは, 日本語に対しては形態素解析システム MeCab[23] の分かち書き, 英語に対しては自然言語処理ツールキット NLTK[24] のトークナイザなどによって文章を単語に分割したのちに, 単語の出現傾向を学習用メール群から求め, それを分類対象のメールが含む単語に当てはめることでメールを分類する.

すなわち, 分類対象のメールに出現する単語の出現傾向が, 学習用メール群のものと似ているほど分類性能が高くなり, 逆に異なるほど分類性能が低くなる.

スパム送信者は, この分類性能の低下を意図的に引き起こすため, 単語の中に, 以下の例のように記号, スペース及び HTML タグなどの組み

込みにより，改変した文字列を作成して用いることがある [25].

- price$ for be$t drug$!
- Sym8oL
- priceC I A L I S
- <font> se</ font> xu< font> al</ font>

このような語は辞書に載っておらず（以後，辞書なし語と呼ぶ），上記の組み合わせの変更によって新しいものが日々作られている．このことが，スパムメールの特徴が時間的変化しやすい原因の一つとなっている [26].

この中でも特に，学習用メールが含まない新しい単語については，学習できていないため，分類に利用できない（以後，分類のみ単語と呼ぶ）．

単語の特徴の変化による悪影響を抑えるための手法として，オンライン機械学習を用いた研究 [8] がある．これは，機械学習を逐次的に行うことで，分類のみ単語を減らすことに貢献するが，分類のみ単語を利用するわけではない．

分類のみ単語と辞書なし語を分類に利用する試みは，これまで個別には行われている．

分類のみ単語を利用する手法としては，bsfilter[18] が行っている，大文字と小文字の変換や記号の削除などにより，学習済みデータベース中から一致する単語を探索し，見つかった単語の特徴を当該分類のみ単語に適用する手法があるが，スパム送信者が作成した語の特徴が失われてしまう．

辞書なし語を利用する手法としては，メール本文が含む辞書なし語の数を用いる手法 [28]，辞書なし語と記号の数を用いる手法 [29]，及び一般テキスト処理で行われている，類似語を探索し，その特徴を辞書なし語に適用する手法 [27] などにより，特別な処理を施して分類に利用する手法がある．

その一方で，日本語の新聞記事を分類する研究 [31] では，辞書なし語を未処理のまま利用することで，分類性能が向上することが確かめられている．

これらの研究結果は，辞書なし語の中に，特別な処理を必要とするものと，未処理のまま利用することで分類精度が向上する語の両方を含む

ことを示しているが，これらを区分して扱うための研究は，これまでに行われていないため，本論文ではこれに取り組んだ．

1.2 本研究の目的

辞書なし語の特性解析を行い，その結果を基に，高性能な分類を可能とする辞書なし語の利用手法を開発する．

具体的には，実験に用いるデータセットが含むすべての単語について，学習用と分類用メール群，及び正規とスパムメールに着目した単語の出現パターン区分を定義し，その各々について，分類性能に与える影響や，単語の種類数を調べる．

その結果を，辞書なし語とそれ以外の単語で比較することで，辞書なし語の特性を明らかにする．また，出現パターンごとでの比較も行うことで，辞書なし語を分類のみ語とそれ以外に分けて扱うことによる分類性能向上の可能性を示す．

辞書なし語について，分類のみ語とそれ以外に分け，分類への利用手法をそれぞれ開発する．この効果について，既存フィルタリング手法との併用実験により，分類性能の向上を確かめる．

メールフィルタリングは改良が重ねられ，その性能は限界まで来ている．さらなる精度向上，すなわち完全フィルタリングに近づけるためには，これまでにない観点が必要であり，本論文は辞書なし語の利用という，新しい観点の一つを示すものである．

1.3 本論文の構成

本論文は次のように構成される．

2章では関連研究について説明する．機械学習を用いたフィルタリング手法の種類について，Heuristic or Rule Based，Content Based，Previous Likeness Based，Case Based，及び Adaptive Spam Filtering Technique に分け，それぞれの特徴や，長所と短所について概観する．

この中で，本研究が扱う Content Based Spam Filtering Technique について，その具体的な手法であるベイジアンフィルタ，SVM 及び BONSAI

を用いたフィルタリング手法について説明する．

3章では，実験に必要な準備を行う．具体的には，使用する電子メールのデータセット，本研究が扱う単語の取り扱いと辞書なし語について説明したのちに，分類性能の評価に用いる尺度であるROC曲線，データの集まり具合やばらつき具合を調べるための散布図，及び箱ひげ図について説明を行う．

さらに，単語の出現傾向について詳しく調べるために必要となる，学習用と分類用メール群，正規とスパムメールの違いに着目した単語の出現パターン区分を定義する．

4章では，辞書なし語とそれ以外の単語の特性の違いについて調べる．具体的には，辞書なし語を品詞の一つと捉え，これと名詞，動詞及び形容詞で分類性能の違いを調べるため，これらの各々を用いた分類実験をbsfilterを用いて行い，結果を比較する．これにより，辞書なし語の分類性能が最も高いことが示される．

この原因について調べるため，辞書なし語，名詞，動詞及び形容詞の各々について，単語の種類数を出現パターンに分けて集計して比較する．その結果，第一に，辞書なし語には，正規またはスパムメールの一方にのみ出現し，かつ学習用と分類用メール群の両方に出現する，すなわち最も分類に貢献する出現パターンの単語が多いことを確かめ，これが辞書なし語の分類性能の高さの原因であることを示す．他方で，辞書なし語には，分類性能を低下させる出現パターンの語を名詞，動詞及び形容詞と同程度含むため，このような単語を分類に用いないようにする新たな手法を提案する．

第二に，辞書なし語には，分類用メール群のみに出現する，すなわち学習用メール群になく，学習できない単語が多いことを示す．この単語について，分類に利用するための新たな方法を提案する．

すなわち，辞書なし語を，学習用と分類用のメール群の両方に出現する語と，分類用メール群のみに出現する単語に分けて扱い，その各々で分類への利用手法を新たに提案する．これらの手法は，辞書なし語の特徴をより効率よく分類に利用するため，分類性能が向上できることを次章以降で示す．

5章では，辞書なし語のうち，学習用と分類用メール群の両方に出現する語の利用手法を開発する．具体的には，分類性能を低下させる出現

パターンの語の多くを除外し，分類に大きく貢献する出現パターンの語の多くを分類に用いる手法を開発する．まず，これらの単語を区分できる特徴について考える．

学習用と分類用メール群の両方で，出現傾向が変わらずに出現する単語が分類に貢献するものである．このような単語は長い期間で継続して用いられ，多くのメール本文や件名に出現すると考えられる．一方で，学習用と分類用メール群で出現傾向が大きく変化する単語が分類性能を低下させるものである．このような単語は一時的に用いられるものを含み，出現するメールの数（文書頻度）が少ないと考えられる．

すなわち，文書頻度に着目した解析を行うことで，これらの特徴の違いを見出せると考え，これを試みる．具体的には，出現パターンごとに，各々の単語について文書頻度を集計し，その結果を比較する．これにより，分類性能を低下させる出現パターンの語の文書頻度が，その他の単語よりも低い傾向にあることを示す．

つまり，文書頻度にあるしきい値を定め，これを超える単語のみを分類に用いることで，分類性能を低下させる出現パターンの語の多くを除外できるということである．このしきい値について，最適な値を探るため，文書頻度のしきい値と分類性能の関係について調べる．具体的には，bsfilter を用い，しきい値を変化させながら分類実験を繰り返し行い，その結果を比較する．これにより，しきい値を 1 から 7 付近に変化させるにつれ，分類性能が向上することを示す．

6 章では，辞書なし語のうち，分類用メール群にのみ出現する語（分類のみ単語）の利用手法を開発する．この単語は，学習用メール群から時間が経過するほど出現しやすいと考えられ，このことが分類性能の低下を招くと考えられる．まず，学習用メール群からの時間の経過による分類性能の低下具合について確かめるため，分類用メール群を，受信日時の順番に並べたうえで 8 等分し，さらに正規とスパムメールに分け，その各々で bsfilter を用いた分類実験を行い，結果を比較する．これにより，分類性能の低下は，正規メールではほとんどなく，スパムメールで大きいことを示す．

この原因について調べるため，8 つに分けた分類用メール群ごとに，メール本文や件名に出現する分類のみ単語の種類数を，正規とスパムメールに分けて集計し，その結果を比較する．これにより，辞書なし語の分類

のみ単語が，正規よりもスパムメールに多い傾向にあることを示し，これが先に述べたスパムメールの分類性能の低下の原因であることを示す．

この傾向を分類に利用するため，辞書なし語の分類のみ単語が多く出現するメールを，スパムメールに分類しやすくするようにバイアスをかける手法を提案する．この手法と既存手法の併用方法を，bsfilter，SVM 及び BONSAI を例に挙げ，それぞれ説明する．

本論文では bsfilter との併用について扱い，辞書なし語の分類のみ単語にスパム確率を一律に設定する手法を提案する．設定するスパム確率の最適な値を探索するため，スパム確率を 0.0（最小値）から 1.0（最大値）まで変化させながら分類実験を行い，その各々の結果をオリジナルの bsfilter の結果と比較する．これにより，辞書なし語の分類のみ単語に対し，0.7 付近のスパム確率を一律に設定することで，オリジナルの bsfilte と比べて大きく分類性能が向上することを示す．

7 章では，5 章で開発した学習用と分類用の両方に出現する語の利用手法と，6 章で開発した分類用メール群にのみ出現する語の利用手法を組み合わせた処理の流れについて説明し，今後の展望を述べ，本論文をまとめる．

第2章　関連研究

2.1　機械学習を用いたフィルタリング手法の種類

現在，スパムメールを分類するため，さまざまなフィルタリング手法が提案されている．文献 [5] では，機械学習を用いたフィルタリング手法について概観しており，Heuristic or Rule Based，Content Based，Previous Likeness Based，Case Based，及び Adaptive Spam Filtering Technique に分け，その有効性について述べている．これらについて次で詳しく説明する．

2.1.1　Heuristic or Rule Based Spam Filtering Technique

SpamAssassin[6] などで用いられている手法である．これは，経験則に基づき，人手で作成したスパム検出のためのブラックリスト，または正規メール受信のためのホワイトリストを作成し，これらと新たに受信したメールの情報と照らし合わせることにより，メールを分類する手法である．具体的には，メールアドレス，ドメイン名，及びメールの送信に用いられたメールサーバーの情報などのキーワードを用いることが多い．

この手法は，「1. 1 本研究の背景」で述べた，時間経過とともに特徴が変化するメールに対応するために，リストの更新を定期的に人手で行う手間を必要とする．また，文献 [5] では，近年のスパム送信者のフィルタすり抜け技術の進歩により，この手法が機能しづらくなっていることが指摘されている．

2.1.2 Content Based Spam Filtering Technique

現在最も広く用いられている機械学習に基づく手法であり，その分類性能は完全フィルタリングに近づいている．

これは，学習用メール群から学習した単語の出現傾向や分布を，分類対象のメールに当てはめることで自動的にメールを分類する手法である．代表的なものには，ベイジアンフィルタリング手法 [4]，SVM を用いた手法 [8]，決定木に基づく機械学習システム BONSAI[9] を用いた手法 [11]，及びニューラルネットワークを用いた手法 [12] などがある．

この手法は，単語の特徴を事前に学習しておくため，学習用メールデータの規模に関係なく，メール分類にかかる処理時間が短いという利点があるため，多くのメール利用者が共有する大規模なメールサーバでの運用に適している．

2.1.3 Previous Likeness Based Spam Filtering Technique

K-Nearest Neighbor(k-NN)[13] に代表される手法である．これは，学習用メール群について，メール同士の類似度に基づく位置関係を多次元空間で学習する．メールを分類するとき，学習した位置関係の中に当該分類対象のメールを付置し，その周辺のメール群のラベル（正規またはスパム）の位置関係に基づき，距離が近いラベルに分類する．

これにより，学習用データをすべて用いる Content Based Spam Filtering Technique と比較して，新たな特徴を持つメールに素早く対応できるという利点があるが，学習に用いるメールデータが多くなるにつれ，類似メールの探索に要する処理が増大するため，分類に要する時間が増大するという欠点がある．

2.1.4 Case Based Spam Filtering Technique

K-Nearest Neibor(k-NN) を用いた ECUE[14] に代表される手法である．これは，事前学習を行わず，新しいメールを受信するたびに，それに似

たメールを学習用メール群から探索し，見つかったメールのみでその都度学習を行い，メールを分類する手法である．

この手法は，分類対象のメールと無関係のものを用いないため，局所的な特徴も捉えられ，高性能な分類が期待できる一方で，k-NN などを用いた類似メールの探索，及び分類用メールごとに学習処理を行うため，メール分類に要する処理時間が長いという欠点がある．

2.1.5 Adaptive Spam Filtering Technique

POPFile[15] で用いられている手法である．これは，機械学習に用いるデータセットを，正規またはスパムメールの 2 種類だけでなく，より細かいジャンル分けを行い，その各々で特徴を学習する手法である．

分類手法としては，前述した Content Based Spam Filtering Technique が用いられることが多く，分類対象のメールをジャンルに分類する．

この手法は，学習用メール群を，フィルタ利用者によってあらかじめジャンル分けをしておく手間を必要とする．また，その区分のルールは受信者ごとに異なるため，多くの利用者が共有するメールサーバでの運用は難しい．

2.2 本研究で扱うフィルタリング手法

本研究では，現在主流である Content Based Spam Filtering Technique を扱う．具体的には，現在需要が高い [2] とされるベイジアンフィルタを主体とし，SVM，決定木に基づくフィルタリング手法も扱う．これらの手法について次で詳しく説明する．

2.2.1 Paul Graham 方式のベイジアンスパムフィルタ [3]

ベイズ理論に基づき，統計的にスパムメールを分類する手法である．

過去に受信したメール群（学習用メール群）を正規とスパムメールに分け，件名や本文に出現する各単語について，出現頻度を正規とスパムメールのそれぞれで集計し，単語 w_i のスパム確率 $p(w_i)$ を次式で求める．

$$p(w_i) = \frac{\frac{b_i}{n_{bad}}}{\frac{g_i}{n_{good}} + \frac{b_i}{n_{bad}}} \tag{2.1}$$

ここで，b_i（g_i）は，スパム（正規）に出現する単語 w_i の頻度であり，n_{bad}（n_{good}）は，スパム（正規）メールの総数である．

この方式では，開発者の経験則的に，$b_i + g_i$ が 4 以下の単語を学習用メール群から除外するのがよいとされ，さらに，分類対象のメールに初めて出現する単語のスパム確率を 0.4 とすることで，分類性能が向上するとされている．

単語のスパム確率 $p(w_i)$ を用い，メールのスパム確率 $p(d)$ を次式で求める．

$$p(d) = \frac{\prod_{i=1}^{n} p(w_i)}{\prod_{i=1}^{n} p(w_i) + \prod_{i=1}^{n} (1 - p(w_i))}$$

ただし，この計算には，特徴的な単語のみを扱うため，スパム確率が 0.5 から最も離れている 15 個の単語を用いる．

2.2.2 Gary Robinson 方式のベイジアンスパムフィルタ [16]

Paul Graham 方式を改良した手法であり，主に出現頻度が低い単語の扱い方が改善されている．

単語のスパム確率 $p_{GR}(w_i)$ を，式 2.1 で求めた $p(w_i)$ を用い，次式で求める．

$$p_{GR}(w_i) = \frac{s \cdot x + n_i \cdot p(w_i)}{s + n_i} \tag{2.2}$$

ここで，x は分類対象に初めて出現する単語のスパム確率であり，n_i は学習用メール群での単語 w_i の出現頻度である．s は，出現頻度（n_i）が少ない単語ほど，そのスパム確率を x に近づけるためのバイアスであり，大きいほど，低頻度の単語のスパム確率が x に近づく．

式 2.2 で求めた $p_{GR}(w_i)$ を用い，メールのスパム確率 $p_{GR}(d)$ を次式で求める．

$$P = 1 - \left(\prod_{i=1}^{n}(1 - p_{GR}(w_i))\right)^{\frac{1}{n}} \tag{2.3}$$

$$Q = 1 - \left(\prod_{i=1}^{n} p_{GR}(w_i)\right)^{\frac{1}{n}} \tag{2.4}$$

$$p_{GR}(d) = \frac{P-Q}{P+Q}$$

2.2.3 Gary Robinson-Fisher方式のベイジアンスパムフィルタ [17]

高性能な分類が可能となっており，POPFile[15]や bsfilter[18] などで広く用いられている．

この手法は，単語の出現確率がそれぞれ独立であると仮定し，Gary Robinson方式の式 2.2で求めた単語のスパム確率 $p_{GR}(w_i)$ を掛け合わせた次式の値を用いる．

$$\prod_{i=1}^{n} p_{GR}(w_i) \tag{2.5}$$

さらに，すべての単語の出現傾向が一様分布（正規とスパムに関係なく一様に出現）であるという仮説のもとに，検定を行う．具体的には，式 2.5で求めた値を用い，確率変数 x を次式で求める．

$$x = -2\ln \prod_{i=1}^{n} p_{GR}(w_i) \tag{2.6}$$

これが自由度 2n のカイ二乗分布に従うと考える．例えば，自由度 10 のときのカイ二乗分布のグラフは，図 2.1のようになる．p 値は，式 2.6で求めた確率変数 x 以上の領域（赤枠部）の面積であり，次式のように逆 χ^2 関数（C^{-1}）を用いることで求めることができる．

$$p\text{ 値} = C^{-1}\left(-2\ln \prod_{i=1}^{n} p_{GR}(w_i); 2n\right) \tag{2.7}$$

帰無仮説は「単語がそれぞれ独立であり，かつ一様分布」であり，p 値が小さい（一般的には 0.05 より小さい）ときにこれを棄却する．実際には，単語は特定の組み合わせで用いることもあるため独立でなく，また，

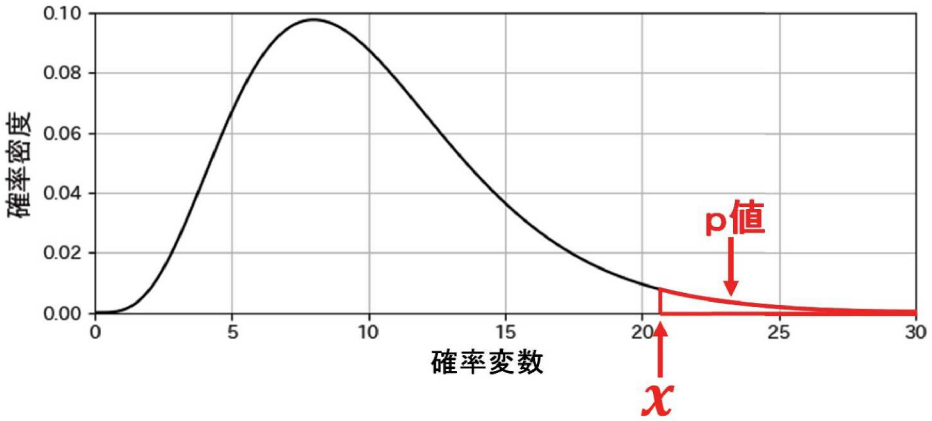

図 2.1: 自由度 10 のときのカイ 2 乗分布と p 値

正規またはスパムメールのどちらかに偏って出現する単語もあることから，帰無仮説が間違っていることがわかる．すなわち，p 値が小さくなることが尤もらしい．

ここで，式 2.7 の p 値が小さくなる（図 2.1 の赤枠部が小さくなる）条件を考えると，式 2.6 の確率変数 x が大きくなるときであるため，式 2.5 の単語のスパム確率の積が小さいとき，すなわちスパム確率の低い単語が多いときである．つまり，正規メールが含む単語で求めた p 値は小さくなる傾向にある．

Gary Robinson-Fisher 方式では，この p 値の小ささを正規メール度 H として，次式で定義している．

$$H = C^{-1}(-2\ln \prod_{i=1}^{n} p_{GR}(w_i); 2n)$$

ここで用いている $p_{GR}(w_i)$ を，1 から引いたものに変えることで，スパムメール度 S についても次式で定義している．

$$S = C^{-1}(-2\ln \prod_{i=1}^{n} 1 - p_{GR}(w_i); 2n)$$

これらの正規メール度 H とスパムメール度 S を用い，メールのスパム確

率 $p_{GRF}(d)$ を次式で求める．

$$p_{GRF}(d) = \frac{1 + H - S}{2}$$

2.2.4 bsfilter[18]

現在広く用いられているフィルタである．分類手法として，Paul Graham提案の方式[3]，Gary Robinson提案の方式[16]，及びGary Robinson-Fisher方式[17]の3種類のベイジアンフィルタリング手法を選択できるが，本研究では最も後発であり，高性能とされているGary Robinson-Fisher方式を用いる．

bsfilterでは，表記ゆれを軽減するため，学習しておらず，分類対象のメールに初めて出現した分類のみ単語の各々に対し，次のような処理を順に施しながら，学習済みデータベース中の単語と照合し，一致した場合はその単語のスパム確率を当該分類のみ単語に適用している．

1. 末尾文字を削除
2. 「j」「!」「*」「'」を削除
3. 全ての文字を小文字に変換する
4. 全ての文字を大文字に変換する
5. 先頭文字のみ大文字にする

すなわち，学習済みデータベースにない分類のみ単語については扱わない．また，辞書あり語と辞書なし語の区別も行っていない．

2.2.5 SVMを用いたフィルタリング手法[20]

サポートベクトルマシン（SVM）を用いたメールフィルタリング手法は，高次元のベクトルを用いてメールを分類する手法である．

具体的には，正規とスパムの2つの学習用データのベクトル集合を，

$$(x_1; y_1); \cdots; (x_l; y_l); x_i \; 2 \; R^n; y_i \in -1; +1 \tag{2.8}$$

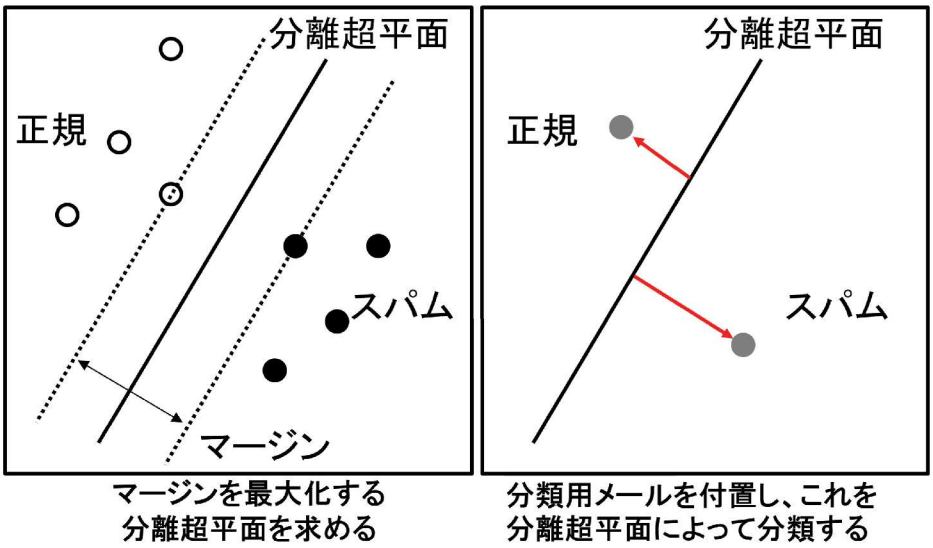

図 2.2: SVM を用いたメールフィルタリング手法

とする．ここで，x_i はメール i の特徴ベクトルで，y_i はメール i がスパム（1）正規（-1）かを表すスカラーである．

メール分類では，メールの特徴を，メール中に出現する単語で代表させ，単語 w_i が出現する場合，$w_i = 1$，出現しない場合を，$w_i = 0$ としてメールをベクトル

$$x_i = (w_1; w_2; \cdots; w_n)$$

で表す．

式 2.8 の多次元ベクトルを用いることで，図 2.2 の左に示すように，学習用メール群を多次元 Euclid 空間上に付置し，これらをスパムと正規に二分する分離超平面を学習する．

この分離超平面は，マージン（分類超平面に最も距離が近い正規とスパムメールとの距離）を最大化するように求める．

具体的には，分離超平面

$$w \cdot x + b = 0 \tag{2.9}$$

を考える．これを正規とスパムメールをそれぞれ，次の領域に分割する．

$$w \cdot x_i + b \geq 1 \text{ if } y_i \ 2 \ 1 \tag{2.10}$$

$$w \cdot x_i + b \leq -1 \text{ if } y_i \in -1 \tag{2.11}$$

式 2.10，式 2.11 をまとめると，

$$y_i(w \cdot x_i + b) \geq 1 \ (i = 1; \cdots ; l) \tag{2.12}$$

となり，式 2.12 に属さない領域がマージンの領域である．ここで式 2.9 の分離超平面と個々の学習メール x_i の距離 $d(w; b; x_i)$ は

$$d(w; b; x_i) = \frac{|w \cdot x_i + b|}{\|w\|} \tag{2.13}$$

となる．マージンは，学習用メールの中で，分離超平面に最も近い正規とスパムメールの間の距離であるため，

$$\min_{x_i(y_i=1)} d(w; b; x_i) + \min_{x_i(y_i=-1)} d(w; b; x_i) \ (i = 1; \cdots ; l)$$
$$= \frac{1}{\|w\|} + \frac{1}{\|w\|} = \frac{2}{\|w\|}$$

となる．これを最大化するためには，$\|w\|^2$ を最小化すればよい．以上のことから，これは次の制約付き最適化問題に定式化できる．

$$\begin{aligned} &\text{minimize} && \|w\|^2 \\ &\text{subject to} && y_i(w \cdot x_i + b) \geq 1 \ (i = 1; \cdots ; l) \end{aligned}$$

この制約付き最適化問題は，Lagrange 乗数を用いると，次のより扱いやすい双対問題に帰着できる．

$$\begin{aligned} &\text{minimize} && L(a) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i;j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \\ &\text{subject to} && \alpha_i \geq 0; \ \sum_{i=1}^{l} \alpha_i y_i = 0 \ (i = 1; \cdots ; l) \end{aligned}$$

この式で $\alpha_i \geq 0$ となる x_i を Support Vector と呼ぶ．w はこれらの Support Vector より

$$w = \sum_{i(ALLSV)} \alpha_i y_i x_i$$

となり，b は

$$b = -(w \cdot x_i + \frac{1}{y_i})$$

で求められる．

メールを分類するためには，図 2.2 の右図に示すように，分類対象のメール x_i と分離超平面との距離 $d(w; b; x_i)$ を式 2.13 によって求めればよい．実際には，正規またはスパムの二項分類であるため，距離の長さについては扱わず，符号のみを用いて分類する．

2.2.6 BONSAI を用いたフィルタリング手法 [11]

機械学習システム BONSAI は，一次元の記号列を対象とし，学習用の正例の群と負例の群のデータベースから，それらを分類する規則を決定木の形式で導き出す [9].

BONSAI を用いたメールフィルタリング手法の全体像を図 2.3 に示す．まず，学習用メール群を，あらかじめスパムと正規に分けて用意し，各々のメールについて，文章を単語に分かち書きする．

次に，各々の単語について，次式によって正規とスパムメールの出現頻度の偏りを求め，これを単語の特徴の強さの指標として用いる．

$$p(w_i) = \frac{b_i}{g_i + b_i}$$

ここで，b_i (g_i) は w_i が学習用メール群のスパム（正規）に出現する総数である．この値と表 2.1 に示す記号への変換ルールを照らし合わせ，単語を記号へ置き換える．

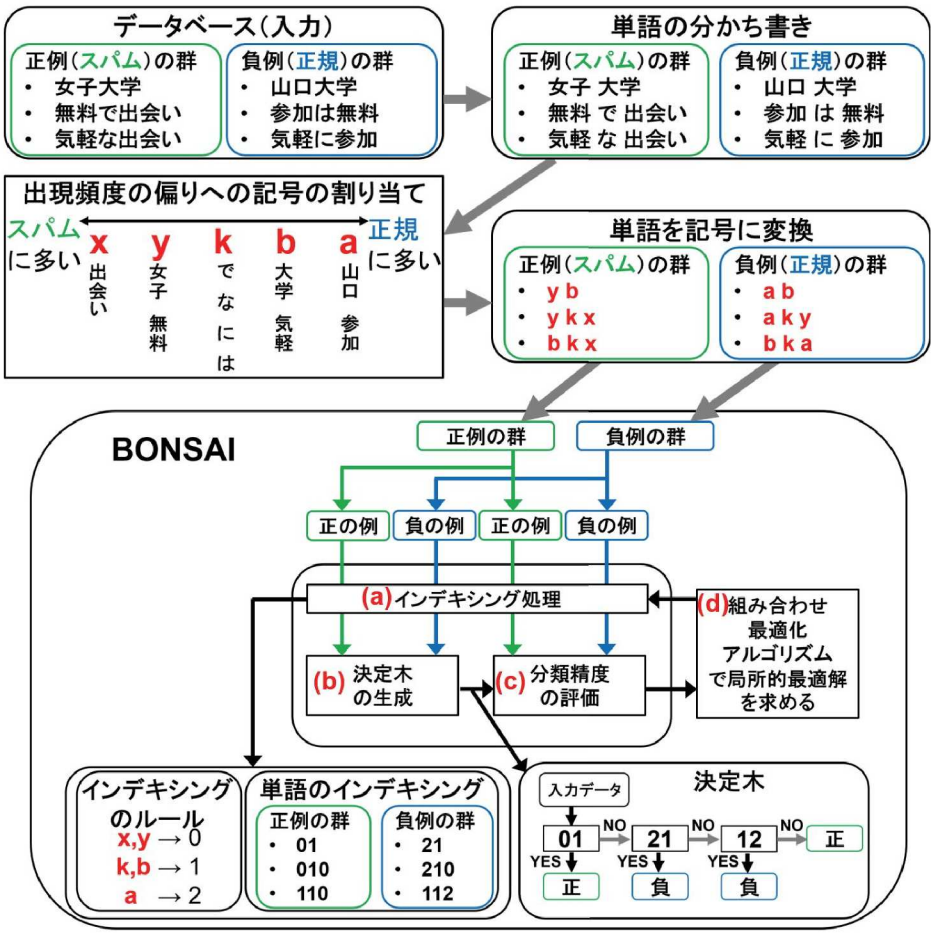

図 2.3: BONSAI を用いたメールフィルタリングの流れ (文献 [11] から引用し一部改変)

これにより，メールデータが含むすべての単語を，正規とスパムメールの出現傾向の偏りを表す「x,y,k,b,a」の 5 種類の記号に置き換えることができ，1 通のメールが含む単語群を一次元の記号列，すなわち BONSAI が対象とする形式に変換することができる．

次に，一次元の記号列に変換した正例と負例の両方のデータ群を BONSAI に入力する．BONSAI では，正例と負例をそれぞれ読み込み，次の

表 2.1: 単語の出現頻度の偏りに基づく記号への変換ルール（文献 [11] から引用）

変換記号	単語の出現頻度の偏り
x	0.8 以上
y	0.8 未満，0.6 以上
k	0.6 未満，0.4 以上
b	0.4 未満，0.2 以上
a	0.2 未満

(a)〜(d) の処理を行う．

(a)：学習例が含む記号の並びの規則性から，特徴が似ている記号をインデキシングによってまとめる．このように少ない要素で並びの規則性を捉えることで，より構造的でかつ汎用性のある分類規則を導くことができる．また，要素の種類数が減少することにより，分類規則の生成に要する計算時間を大幅に短縮することもできる．

図 2.3 の例では，BONSAI が「x，y」を 0,「k，b」を 1,「a」を 2 にインデキシングしている．

(b)：インデキシング後の記号列をノードに配置し，正例と負例を最も高性能に分類できる規則を示す決定木を生成する．分類方法については，分類対象の記号列に対して，各ノードに対応する文字列パターンの有無を，始点から終点まで順に確かめていき，到着した正または負に分類する．

図 2.3 の例の BONSAI が出力した決定木を見ると，負（正規）に分類するルールは,「01」がなくてかつ,「21」または「12」を含むメールである．これ以外については正（スパム）に分類する．

(c)：インデキシング及び決定木の生成に用いていない新しいデータを読み込み，それを用いて決定木の分類精度を実際に確かめ，記憶しておく．

(d)：新しいデータを読み込み，再びインデキシング (a)，決定木の出力 (b)，分類精度の評価 (c) を行い，決定木とその分類精度を確かめる．その後，新しく評価した分類精度と，前回に記憶していた分類精度を比較し，分類精度が高い方の結果を採用し，それを記憶する．

以後（d）の最適化アルゴリズムを繰り返していくことで，より高性能な分類ができる決定木を導き出せる．

BONSAI を用いたフィルタリング手法では，インデキシングによって特徴が似ている単語をまとめたうえで，その並び順から分類規則を導き出すため，正規またはスパムメールの特徴の違いを構造的に捉えた汎用性の高い分類性能を持つ．

第3章　準備

3.1　使用した電子メールデータセット

第三者による実験の再現，及び比較を可能とするため，公開されている次の2つのデータセットを用いた．

SpamAssassin public corpus[32]
: メールフィルタリングの研究[33]で広く用いられているデータセットである（以後，SpamAssassinと呼ぶ）．これは，2002年1月から2003年12月までの約2年間で受信したメール（正規メール4,150通，スパムメール1,987通，合計6,173通）で構成されており，スパムトラップ（スパムメールを受信するためのおとりメールアカウント）から受信したメールは含んでいない．

TREC 2007 spam corpus[34][35]
: Text REtrieval Conferenceで使用され，メールフィルタリングの研究[36][37]で広く用いられているデータセットである（以後，TRECと呼ぶ）．

  これは，2007年4月8日から7月6日までの約3カ月間に，特定のサーバーで受信した電子メール（正規メール25,220通，スパムメール50,199通，合計75,419通）で構成されており，スパムトラップが受信したメールも含んでいる．

本研究で扱うメールの情報は，受信者が実際に必要とする情報に限定するために以下の4つとし，ヘッダ情報は対象外とした．

- メール本文

- 送信者アドレス
- 受信者アドレス
- 件名

3.2 本研究が扱う単語の取り扱いと辞書なし語

テキストデータから単語を抽出するためには，分かち書きを行う必要がある．英語については，単語がスペースによって区切られているため，これを目印に分割することで，機械的に分かち書きができる．他方で，日本語のように単語を繋げて表記する言語については，Mecab[23] などの形態素解析システムを用いる必要がある．

例えば，図 3.1 は，Python 言語により，形態素解析 Mecab を用いて「すももももももものうち」を処理した結果である．このように，分かち書きだけでなく，品詞の推定も同時に行えている．

3.1 で述べた本研究が扱うメールデータは，英語が主であるため，これに対応できる自然言語処理ツールキットである NLTK[24] を用い，トークナイザの機能によって分かち書きをしたものを単語とする．

これによる分かち書きは，単語の末尾の「,」や「.」などの削除といった，単語の整形が自動的に行われるため，スペース区切りによって機械的に分かち書きしたものよりも扱いやすい．

3.2.1 スパムメールが含む辞書なし語

メールフィルタは，過去に受信したメール群から単語の出現傾向を学習し，正規またはスパムメールに特徴的な単語を見出し，これを手がかりとしてメールを分類する．

図 3.2 に示すように,「高額当選」を含むメールを受信したとき，スパムの特徴としてこの単語を学習済みであるため，スパムメールの確率が高いとフィルタが判断し，正しく分類できる．逆に，過去に受信したことのない単語で構成されたメールについては，学習していないため，誤分類しやすい．スパム送信者は，この誤分類を意図的に引き起こすため，

```
wakati.parse("すももももももものうち").split()
```

```
['すもも',
 '名詞,一般,*,*,*,*,すもも,スモモ,スモモ',
 'も',
 '助詞,係助詞,*,*,*,*,も,モ,モ',
 'もも',
 '名詞,一般,*,*,*,*,もも,モモ,モモ',
 'も',
 '助詞,係助詞,*,*,*,*,も,モ,モ',
 'もも',
 '名詞,一般,*,*,*,*,もも,モモ,モモ',
 'の',
 '助詞,連体化,*,*,*,*,の,ノ,ノ',
 'うち',
 '名詞,非自立,副詞可能,*,*,*,うち,ウチ,ウチ',
 'EOS']
```

図 3.1: Mecab の実行例（Python 言語）

「高額当選」などの過去に用いた特徴的な単語について，悪意をもった改変を行い，新しいものに作り変えることがある．

具体的には，以下のような改変があり，TREC データセットが実際に含んでいた単語を例として挙げる．

- 単語中に HTML タグ，記号またはスペースなどを挿入する改変（例：「C I A L I S」,「 se xu al」など）
- アルファベットの追加や削除による改変（例:「Ciialis」,「Vigra」,「Viiagra」など）
- アルファベットを，それに似た記号や数字などで置き換える改変（例：「Cial1s」,「price$ for be$t drug$!」,「Sym8oL」など）

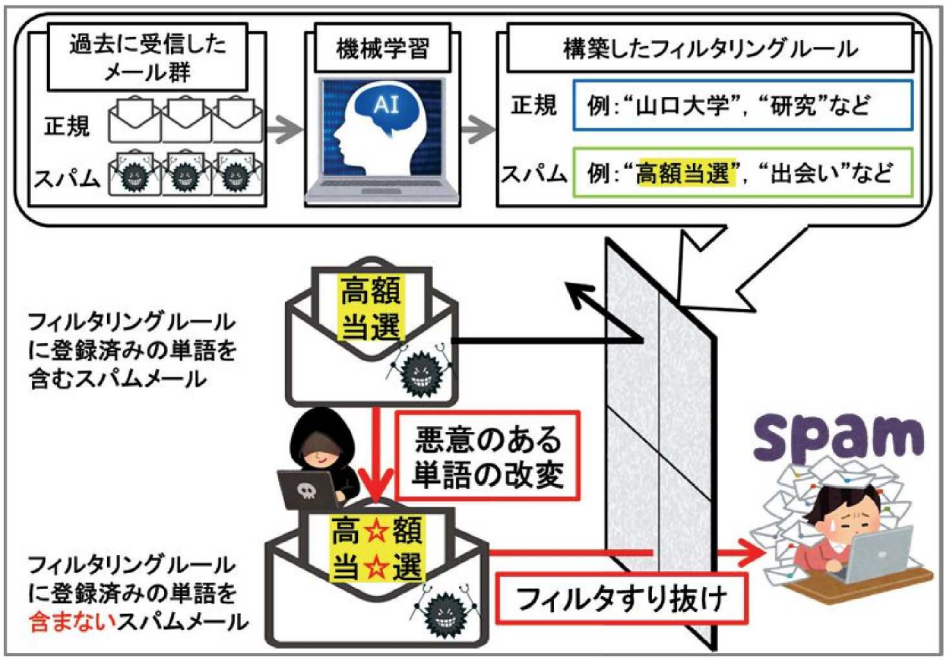

図 3.2: 辞書なし語利用によるフィルタすり抜け

- 英単語の区切りであるスペースを削除することで，単語を結合する改変（例：「ChanceViagra10」，「CialisAs」，「ClickHere」など）

これらの改変により作られた単語は，辞書なし語となる．

本研究では，メール本文や件名に出現する単語について，既存の単語辞書と照らし合わせることで，登録されていないものを見つけ出し，それを辞書なし語とする．この探索をより正確に行うためには，登録されている単語数が多い辞書を用いた方がよいため，データベースに登録されている語数が約 15 万と大規模であり，公開されている WordNet[38] を用いた．

スパムメールが含む辞書なし語については，前で述べた改変によって作り出した語の他に，以下のようなものもある（TREC が実際に含む単語を例に挙げる）．

- 文字化けしている語（例は Tex で表現できないため省略）
- 英語以外の語（例：「もちろんアドレスや電話番号の交換も OK ですよ♪」など）
- URL （例：「/ / ctmay.com」など）
- 数値（例：「15.5」,「14:21:20」など）
- 記号文字（例：「!」,「$」,「%」など）

辞書なし語は，スパムメールのみでなく，正規メールにも出現し，その典型的な例は次のようなものである．

- URL（例:「/ / www.cbsnews.com/ stories/ 2007/ 06/ 20/ health/ webmd/ main2956095.shtml」など）
- 数値（例：「0565」,「03:20:14」など）
- 記号文字（例：「!」,「$」,「%」など）
- プログラムの内容と考えられる単語（例：「strcat」,「strcmp」,「include/config.h」など）

この他にも，正規メールには，新語，親しい間柄で用いる固有名詞，及び略語などが一般に出現する．

3.3 ROC 曲線

メールフィルタリングは，多数の受信メールを，正規またはスパムメールに分類する．このような二項分類は，図 3.3 のように，陰性（例えば正規）の確率密度関数 $!_1$ と陽性（例えばスパム）の確率密度関数 $!_2$ があるとき，これらを特徴量（例えばスパム確率）x のしきい値 x^* によって 2 分するものである．

ここで，しきい値 x^* を超えるものを陽性とすると，次の 4 つの確率を求めることができる [7].

- $P(x < x^* jx\ 2\ !_1)$：真陰性率（$!_1$ が描く面積のうち，x^* より左の面積（青部）が占める割合）

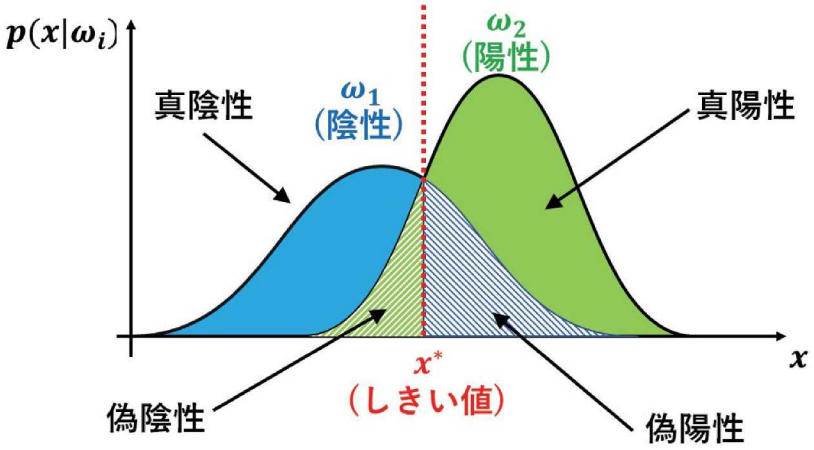

図 3.3: 真陽性，偽陰性，真陰性及び偽陰性の例（文献 [7] から引用し一部改変）

- P(x > x*jx 2 ! 2) : 真陽性率（! 2 が描く面積のうち，x* より右の面積（緑部）が占める割合）
- P(x < x*jx 2 ! 2) : 偽陰性率（! 2 が描く面積のうち，x* より左の面積（緑斜線部）が占める割合）
- P(x > x*jx 2 ! 1) : 偽陽性率（! 1 が描く面積のうち，x* より右の面積（青斜線部）が占める割合）

真陽性率と真陰性率は大きいほどよく，偽陽性率と偽陰性率は小さいほどよい．これらのバランスはしきい値 x* によって変わり，例えば，x* を低く（左に移動）すると，真陽性率は大きくなるが，偽陽性率も大きくなってしまう．

この真陽性率と偽陽性率の関係を，しきい値 x* を最小値から最大値まで変化させながら求めていく．メールフィルタリングを例にすると，図 3.4 の左に示すように，受信メールをスパム確率の順に並べたのちに，メールのスパム確率のしきい値を最大値（1.0）から最小値（0.0）まで変化させながら，真陽性率（スパムを正しく分類する割合）と偽陽性率（正規を誤って分類する割合）を求め，図 3.4 の右の 2 次元のグラフ上にプロットする点の座標（偽陽性率，真陽性率）を求める．

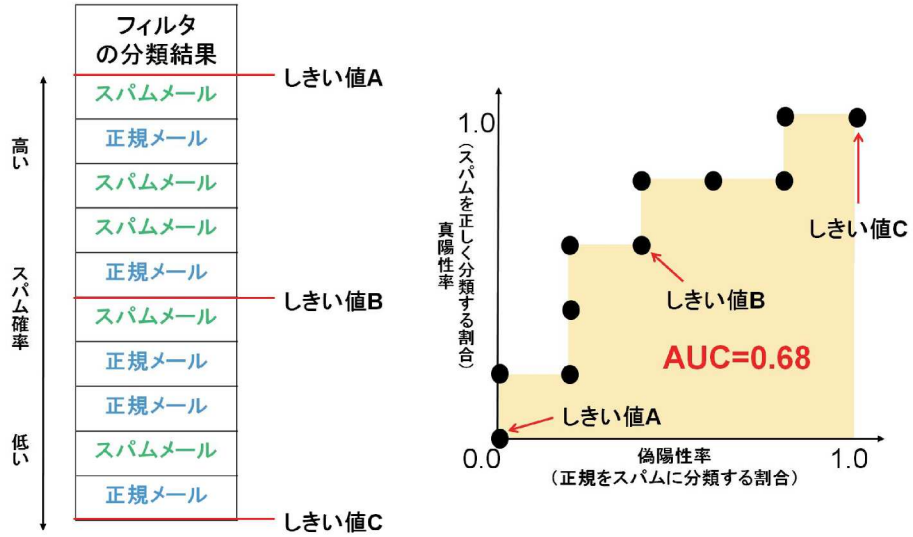

図 3.4: ROC 曲線の説明

例えば，しきい値 A のとき，すべてが陰性（正規）となるため，点（0.0, 0.0）となり，逆にしきい値 C ではすべてが陽性（スパム）となるため，点（1.0, 1.0）となる．しきい値 B のときは，スパムメールを 5 通のうち 3 通正しく分類できているが，正規メールを 5 通のうち 2 通誤って分類しているため，点（0.4, 0.6）となる．このようにしてプロットした点を線で繋いだものが，ROC 曲線 [7] であり，これを分類精度の評価に用いる．

もし，すべてのメールを正しく分類できるフィルタがあったとき，その分類結果は，あるしきい値において，正規とスパムメールを完全に分離できるということである．このとき，真陽性率が最大（1.0），かつ偽陽性率が最小（0.0）となるため，左上隅に点が付置される．すなわち，ROC 曲線が左上隅にあるほど分類精度が高いということである．

この分類精度を数値的に求めるためには，ROC 曲線が描く面積（AUC）を用いればよく，図 3.4 の右では AUC= 0.68 となり，この値が 1.0 に近づくほど分類精度が高い．

このように，ROC 曲線に基づく評価尺度は，しきい値の設定を必要とせずに分類精度を評価できるため，異なる実験環境，あるいはフィルタ

リング手法での実験結果の比較も行える．

3.4 Letter value plots

メールや単語のデータ群は，出現頻度やスパム確率といった尺度で定量化でき，この値はデータごとに異なっていると考えられる．また，この分布についても，正規とスパムメール，メール群といった違いによって異なっていると考えられる．

本研究では，フィルタの分類性能の評価だけでなく，データの分布についても着目して解析する．例えば，図 3.5 の上に示すように，標準正規分布（平均が 0，分散が 1 である正規分布）に従うデータ群を考える．

このデータ群をある程度多く（今回は 10,000 個）用意し，分布を表したのが図 3.5 の散布図である．これを見ると，± 3 以上の領域については，点のばらつき具合を見てとれる一方で，0（真ん中）付近の領域では，点が多く，塗りつぶされているため，分布の形に関する情報が失われている．

図 3.5 の下は，Letter value plots[41] と呼ばれるグラフであり，四分位数を超える分位数も表示した箱ひげ図である．具体的には，最も縦長な箱（青枠部）には，全体のうち 25%のデータがあり，2 番目に縦長な箱（赤枠部）には，全体のうち 12.5%のデータがある．この図は，散布図とは逆に，データが多い箇所を見ることに役立つが，± 3 以上の領域については，点が重なっているため，データの具体的な数を読み取りづらい．

本研究では，これらを組み合わせて，分布について解析する．具体的には，データが少ない箇所を見るために散布図を用い，データが多い箇所を見るために Letter value plots を用いる．

3.5 単語の出現パターン区分

メールの件名や本文に出現する単語は，やり取りをする人との関係，環境及び話題などによって変化することから，単語の出現傾向も，時間の経過と共に変化すると考えられる．

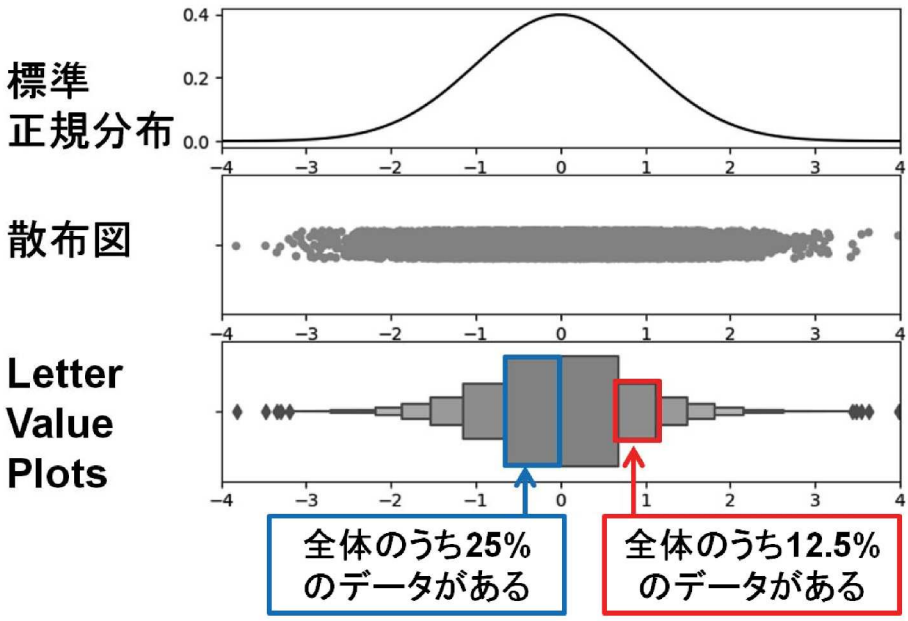

図 3.5: 正規分布（上）に従うデータ群を，散布図（真ん中）及び Letter value plots（下）で描画したグラフ

機械学習に基づくメールフィルタリングは，過去に受信したメール群から単語の出現傾向を学習し，これを新しく受信したメールに当てはめることでメールを分類するため，単語の出現傾向が変化するほど分類性能は低くなる．

スパム送信者は，フィルタの分類性能の低下を意図的に起こすため，新たな辞書なし語を日々作成して用いることにより，単語の出現傾向を変化させることがある．

このように，スパムメールの件名や本文に出現する単語の出現傾向は特に変化しやすく [26]，今後，これに対応できる手法の開発が，フィルタリング性能のさらなる向上のためには必要であるとされている [2]．

本研究では，単語の出現傾向の時間的変化について着目した解析を行うため，単語（w）の出現パターンを次の (a) から (e) の 5 種類に区分して扱う．

(a) :図 3.6 のように，単語 w は，学習用と分類用の両方のメールにおいて，正規（スパム）メールのみが含むという相関が最も強い単語であり，これらが分類に最も役に立つ（以降，最重要単語と呼ぶ）.

(b) :図 3.7 のように，単語 w は，学習用メールでは正規（スパム）メールのみが含んでいたのにもかかわらず，分類用メールではスパム（正規）メールのみが含むという逆の相関を持つ単語であるため，分類用のスパムメールを正規メールに誤分類させてしまう（以降，精度低下単語と呼ぶ）.

(c) :図 3.8 のように，単語 w は，学習用メールのみが含む単語である．このような単語は，学習はできるが，分類用メールに出現しないため，分類のための処理に扱われない（以降，学習のみ単語と呼ぶ）.

(d) :図 3.9 のように，単語 w は，分類用メールのみが含む単語である．このような単語は，学習用メールが含まないため学習できず，特別な処理を施さなければ分類に利用できない．bsfilter では，2.2.4 で述べた 1.〜5. の処理を行うことで，分類のみ単語の一部を分類に利用している.（以降，分類のみ単語と呼ぶ）.

(e) :図 3.10 のように，学習用，分類用，正規またはスパムに関係なく出現する単語であり，このような単語は電子メールが一般に含むありふれた単語が多い.（以降，その他と呼ぶ）.

（a）の出現パターンには，スパム送信者が知り得ない個人名，企業名及び親しい間柄で用いる略語や固有名詞といった，正規メールのみが含む辞書なし語がある一方で，スパム送信者がフィルタをすり抜けるためにわざと作り出した，正規メール送信者が用い得ない辞書なし語も含む.いずれも，正規メールとスパムメールの分類に大きく貢献する重要な単語群である.

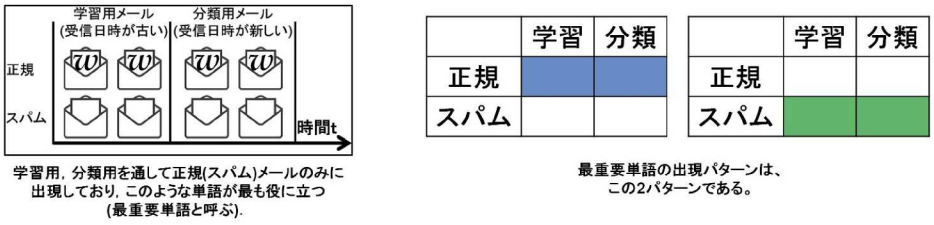

図 3.6: (a) 最重要単語の出現パターン

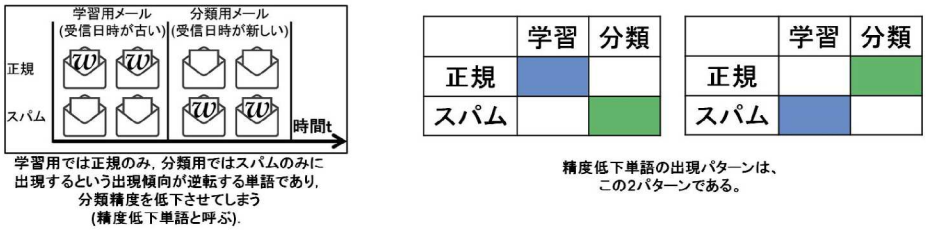

図 3.7: (b) 精度低下単語の出現パターン

図 3.8: (c) 学習のみ単語の出現パターン

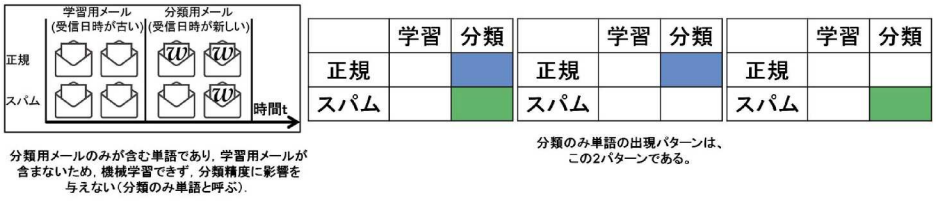

図 3.9: (d) 分類のみ単語の出現パターン

図 3.10: (e) その他の出現パターン

（b），（c），（d）の単語は，分類対象のメールが含む単語の出現傾向が，学習に用いたメール群から変化している．これらの単語が少ないほど，誤分類の可能性は低くなる．

（e）の出現パターンは，メールをやり取りをする人との関係，環境及び話題などに関係なく広く用いられる単語を多く含む．このような単語は，正規とスパムメールの両方が含むことが多く，分類のための決定的な手がかりとはなりづらいが，多くにメールに出現するため，俯瞰的な単語の出現傾向を捉えられるため，メール分類のための手がかりとして貢献する．

第4章　辞書なし語の特性解析と分類への適用

3.2.1で述べたように，辞書なし語には，単語の文字列の中に，記号，HTML，スペースなどを挿入することなどにより，スパム送信者が悪意をもって作成した語を含む．

このような単語は，日常的に用いられている辞書に載っている語とは特性が異なっていると考えられるため，このことについて調査する．

具体的には，辞書なし語を一つの品詞であると捉え，名詞，動詞及び形容詞を比較対象として，分類性能に与える影響，及び3.5で述べた出現パターンを用いた出現傾向の違いについて調べる．

4.1　辞書なし語の分類精度

3.1で説明したTRECデータセットを用い，図4.1に示すように，メール群1，メール群2，及びメール群3として，正規とスパムメールのそれぞれについて，5月1日，5月23日，及び6月14日を基準に，基準日以前の5,000通を学習用，基準日以降の5,000通を分類用メールとして用意した．

用意した正規とスパムメール群はそれぞれ，メール群1〜3を合計すると3万通となるが，3.1で述べたように，TRECデータセットの正規メールの数は25,220通と3万通に満たないため，図4.1に示すように，メール群の区分が重なっている．

すなわち，同じメールが，隣り合うメール群に共に含まれる場合があるが，メール群の重なりは，基準日を超えていないため，学習用メール群同士，または分類用メール群同士でのメールの重複はない．なお，ス

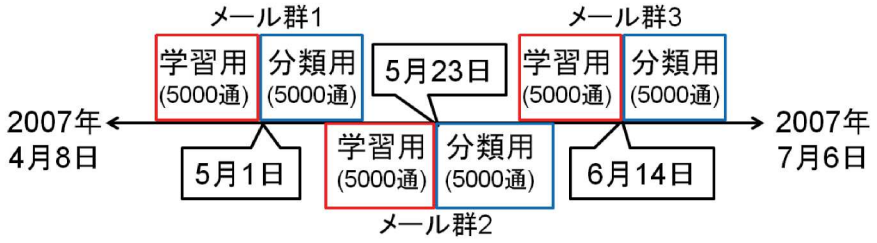

図 4.1: メール群 1，メール群 2，及びメール群 3 の区分

パムメールについては 50,199 通と 3 万通よりも多いため，重複なくメール群が分かれている．

メール群 1〜3 を用い，辞書なし語，名詞，動詞及び形容詞の各々について，bsfilter による分類実験を行った結果の ROC 曲線が図 4.2 である．

これを見ると，3 つのメール群すべてにおいて，辞書なし語の曲線が最も左上隅に近いため，分類精度が最も高いことがわかる．

品詞の区分をなくした場合の結果についても調べるため，辞書なし語のみ，辞書なし語以外の単語，及びすべての単語で同様の分類実験を行った．その結果の ROC 曲線を図 4.3 に示す．この図は，3 つのメール群すべてにおいて，辞書なし語のみでの分類が，すべての単語を用いた分類と同程度に高精度であること，及び辞書なし語以外で分類すると分類精度が低下することを示している．

上記の結果から，品詞ごとの比較，及び辞書なし語以外との比較の両方において，辞書なし語が分類性能に与える影響が大きいことがわかる．

このことから，辞書なし語の扱い方について，特性をよりうまく捉えた分類への利用手法を開発することで，既存フィルタリング手法のさらなる性能向上に貢献すると期待できる．

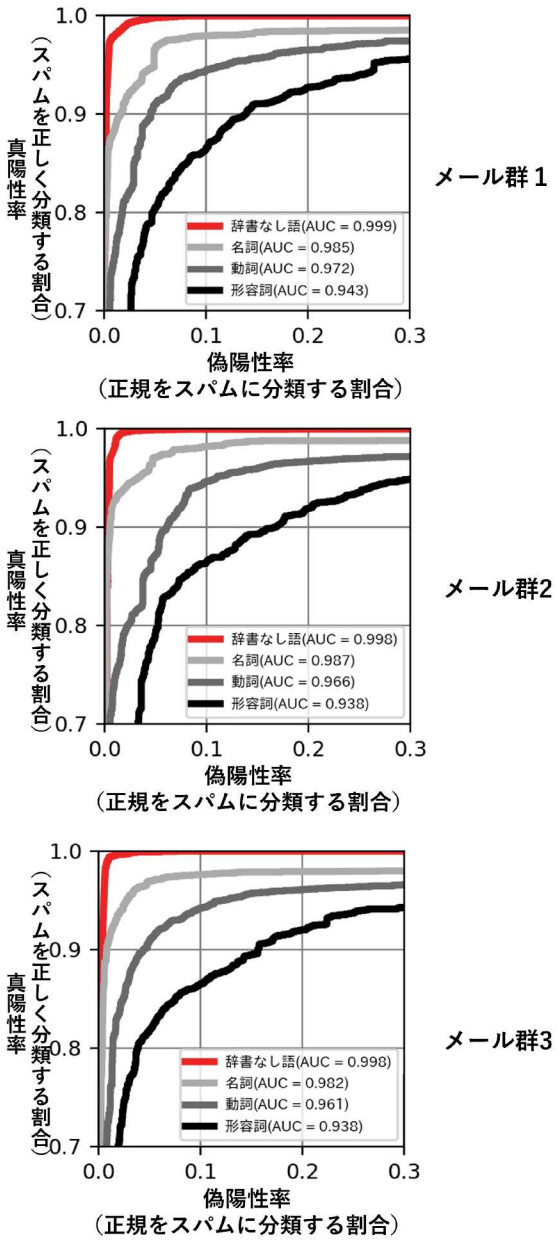

図 4.2: 辞書なし語，名詞，動詞，及び形容詞を用いた分類精度の比較（擬陽性率については 0.3 以下，真陽性率については 0.7 以上の部分を拡大）

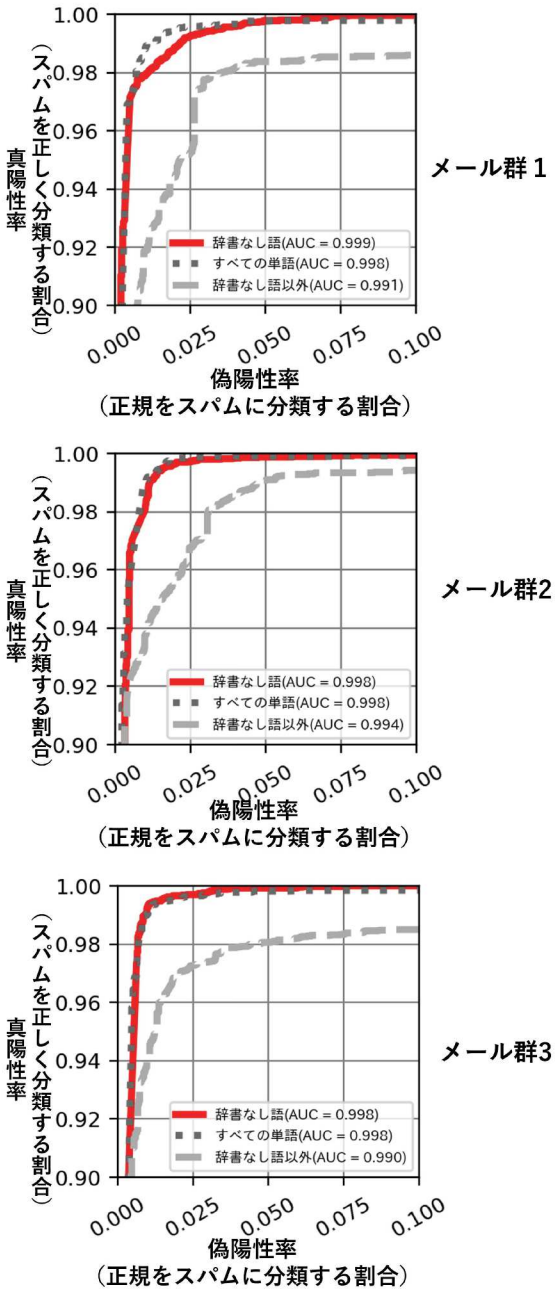

図 4.3: 辞書なし語，すべての単語，及び辞書なし語以外を用いた分類精度の比較（擬陽性率については 0.1 以下，真陽性率については 0.9 以上の部分を拡大）

4.2 出現パターンごとに異なる語の種類数

辞書なし語が分類に与える影響が大きいことを確認できたため，この原因について調べる．

メールフィルタリングは，学習用メール群から捉えた単語の出現傾向を，分類対象のメールの件名や本文に出現する含む単語に当てはめることでメールを分類する．そのため，分類性能の良し悪しは，学習用と分類用メール群との単語の出現傾向の違いに依存し，出現傾向の変化が大きいほど分類性能は低下する．

品詞ごとに，単語の出現傾向の変化について調べるため，辞書なし語，名詞，動詞及び形容詞の 4 種類の品詞それぞれにおいて，4.1 で述べたメール群 1，メール群 2，及びメール群 3 を用い，3.5 で述べた（a）～（e）の出現パターン，すなわち最重要単語，精度低下単語，学習のみ単語，分類のみ単語及びその他の各々について単語の種類数を集計した結果を表 4.1 に示し，それらの比率を棒グラフとして示したものが図 4.4 である．

表 4.1 で辞書なし語，名詞，動詞及び形容詞の合計の単語数を比較してみると，メール群 1～3 のすべてにおいて，辞書なし語が突出して多いことがわかる．その内訳について図 4.4 で確認すると，辞書なし語の学習のみ単語と分類のみ単語の割合（緑と黄色の棒の合計）は，他のものと比較してかなり大きく，合計で約 8 割を占める．

学習のみ単語は分類用メールに出現しないため，分類では扱われない．分類のみ単語は，3.5 で述べたように，分類に利用するためには特別な処理を施す必要がある．

学習用と分類用メール群の両方に出現する単語に着目するために，図 4.4 から学習のみ単語と分類のみ単語を除き，作り直したものが図 4.5 である．これを見ると，精度低下単語については，4 種類の品詞の結果でほぼ同程度に割合が低いこと，及び最重要単語については，辞書なし語のみが約 8 割と多く，名詞，動詞及び形容詞では約 3 割程度と少ないことが確認できる．

このことから，辞書なし語の分類精度の高さの原因は，学習用と分類用メール群の両方に出現する単語の中で，最重要単語が多いことによるものであると考えられる．

表 4.1: 単語の辞書なし語・名詞・動詞・形容詞の分類

	メール群 1			
Words	辞書なし語	名詞	動詞	形容詞
(a) 最重要	25,404	6,825	2,284	1,737
(b) 精度低下	1,379	1,644	556	435
(c) 学習のみ	100,795	10,097	3,602	2,867
(d) 分類のみ	93,880	11,437	4,196	3,221
(e) その他	5,836	13,111	4,731	2,781
合計	227,294	43,114	15,369	11,041
	メール群 2			
Words	辞書なし語	名詞	動詞	形容詞
(a) 最重要	27,498	7,062	2,379	1,705
(b) 精度低下	1,291	1,593	565	399
(c) 学習のみ	104,822	11,595	4,198	3,378
(d) 分類のみ	115,095	10,306	3,828	3,037
(e) その他	6,154	13,647	4,817	2,933
合計	254,860	44,203	15,787	11,452
	メール群 3			
Words	辞書なし語	名詞	動詞	形容詞
(a) 最重要	23,082	6,171	2,175	1,565
(b) 精度低下	1,327	1,527	566	412
(c) 学習のみ	112,367	11,228	4,082	3,257
(d) 分類のみ	86,290	9,922	3,467	2,785
(e) その他	5,688	12,336	4,484	2,573
合計	228,754	41,184	14,774	10,592

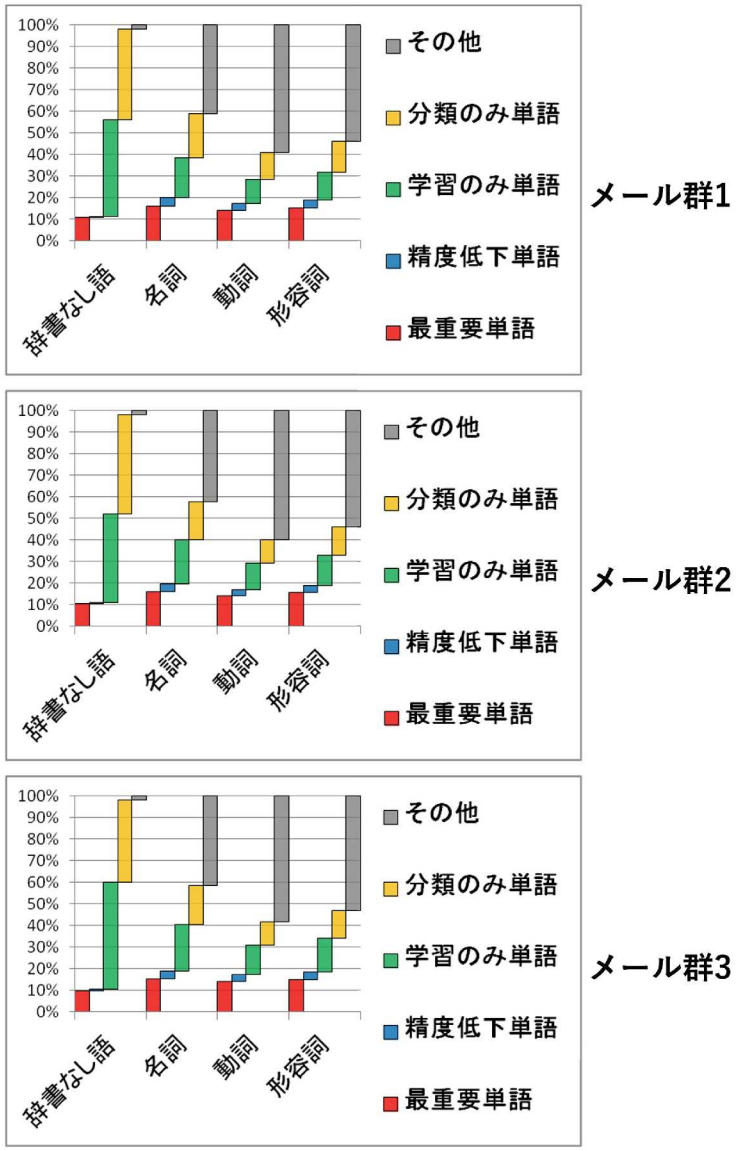

図 4.4: 辞書なし語・名詞・動詞・形容詞における出現パターン別構成比

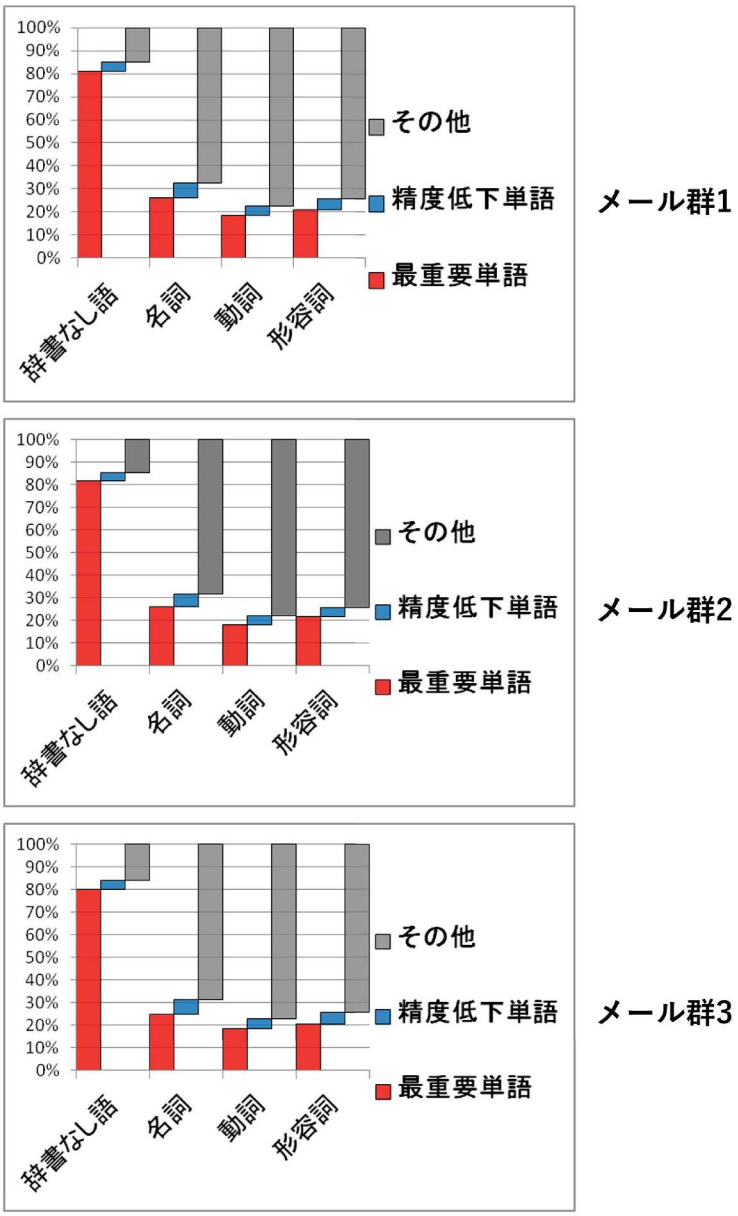

図 4.5: 図 4.4から学習・分類のみ単語を除いた出現パターン別構成比

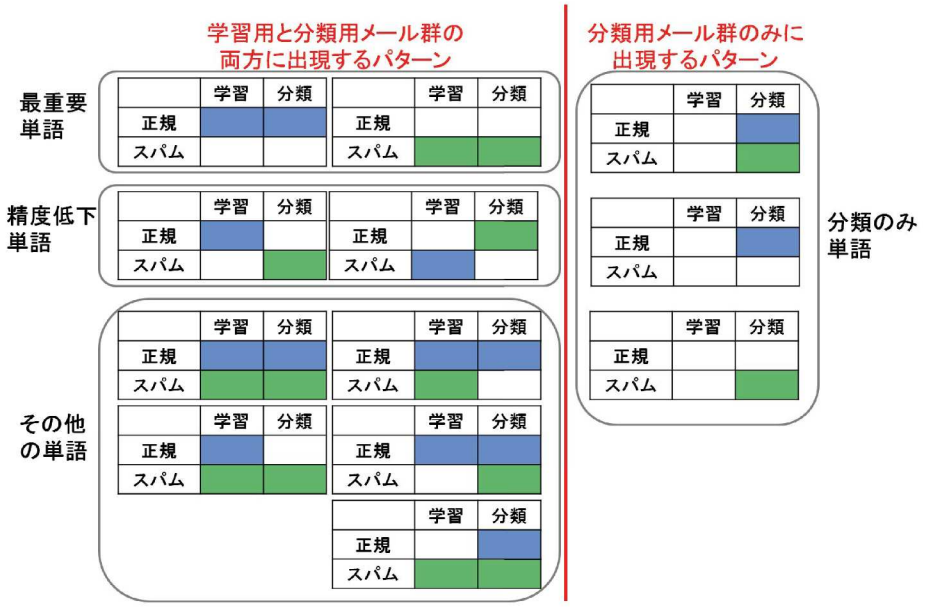

図 4.6: 辞書なし語の利用手法の開発のための区分

4.3 出現パターンに基づく辞書なし語の区分

4.1で，辞書なし語が分類性能に与える影響が大きいことを確認し，その原因について 4.2 で調べたところ，辞書なし語について，学習用と分類用メール群の両方に出現する単語の中で，最重要単語が多いことを確かめた．他方で，辞書なし語全体からみると，学習のみ単語と分類のみ単語がかなり多いことも確かめた．

本研究では，辞書なし語特性をよりうまく分類へ利用する手法の開発に取り組み，既存フィルタリング手法のさらなる分類性能の向上を試みる．

このために，図 4.6 に示すように，辞書なし語を，学習用と分類用メール群の両方に出現する単語と，分類のみ単語に分けて扱い，その各々で利用手法を開発する（学習のみ単語は，分類対象のメールに出現しないため，扱わない）．

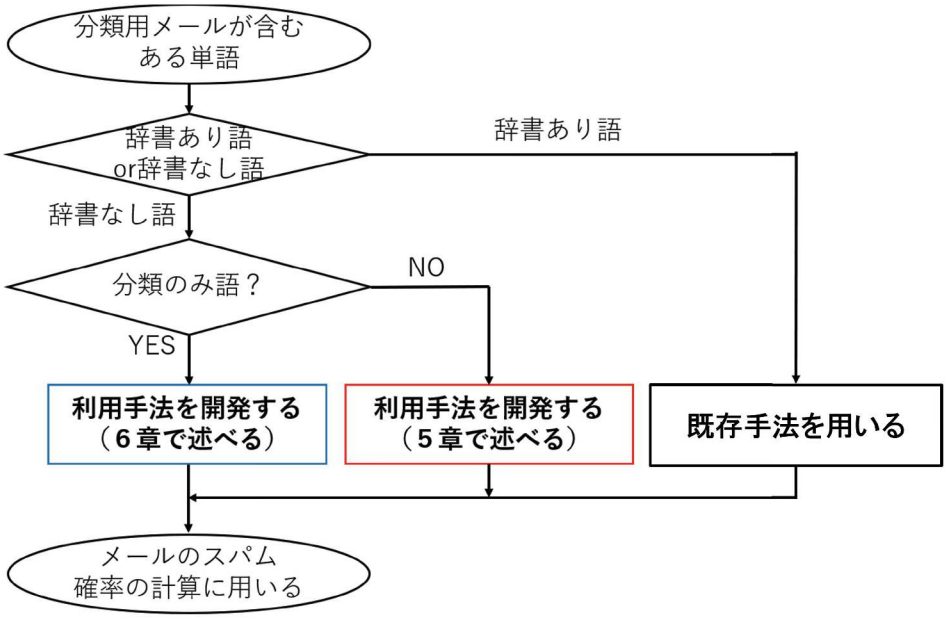

図 4.7: 辞書なし語利用のための処理の概要図

処理の全体像は図 4.7 のようになる．すなわち，分類用メールが含む単語について，辞書あり語と辞書なし語に分岐させ，辞書あり語は既存手法を用いる．辞書なし語については，分類のみ単語と，分類のみ単語以外に分岐させる．

分類のみ単語以外は，学習用と分類用メール群の両方に出現する単語であるが，これは図 4.6 の左に示すように，分類性能の低下を招く精度低下単語も含むため，この多くを分類に用いないようにする手法を開発する．その効果について検証するため，既存フィルタリング手法と併用する実験を行う（この内容については 5 章で述べる）．

bsfilter では，分類のみ単語について，2.2.4 で述べた 1.〜5. の処理を施し，学習用データベースとの照合によって一致した単語として扱うことで，分類のみ単語の一部を分類に利用している．

本研究では，より多くの分類のみ単語を分類に利用するため，辞書なし語の分類のみ単語に着目した特性解析を行い，見出した特性を適用した新たな手法を提案する．その効果について検証するため，既存フィルタ

リング手法と併用する実験を行う（この内容については 6 章で述べる）．

現在高性能な分類が可能となっている既存手法の分類性能を，より完全なものに近づけるためには，フィルタリング手法の改良のみならず，その前段階である単語の扱い方についても改善する新たな観点が必要となる．本論文が提案する辞書なし語の利用は，この観点の一つを与えるものである．

第5章　辞書なし語の文書頻度による分類性能の向上

辞書なし語のうち，3.5で述べた最重要単語，精度低下単語，その他の出現パターンについて，性能向上のための利用手法を開発する．これには，精度低下単語を分類に利用しないようにしつつ，最重要単語の多くを利用するようにすればよいため，これらの特徴の違いについて調べる．

5.1　最重要単語と精度低下単語の文書頻度の違い

最重要単語と精度低下単語の出現パターンはともに，3.5で述べたように，学習用と分類用メール群の両方に出現するパターンであるが，これらの単語の出現傾向が異なる．

具体的には，正規またはスパムメールの片方のみに着目したとき，学習用と分類用の両方に継続して出現するパターンが最重要単語であり，片方にのみ出現するパターンが精度低下単語である

学習用と分類用メールの両方に出現する単語は，ある程度長い期間で継続してメールの件名や本文の中で用いられており，多くのメールに出現していると考えられる．すなわち，精度低下単語と比較すると，最重要単語のほうが，多くのメールに出現する傾向にあると考えられる．

このことについて実験的に確かめるために，辞書なし語の各々について文書頻度を求め，その結果を最重要単語と精度低下単語で比較を行う．

文書頻度 ND(w) は，あるメール d の件名と本文に出現する単語 w の出現頻度を N(w; d) とすると，次式のように表せる．

$$ND(w) = |\{d | N(w; d) \geq 1\}|$$

4.1で述べたメール群 1，メール群 2及びメール群 3を用い，最重要単語，精度低下単語及びその他の各々について，辞書なし語ごとに調べた文書頻度の散布図が図 5.1である．横軸は最重要単語，精度低下単語及びその他を並べている．縦軸は文書頻度であるが，単語ごとに文書頻度が大きく異なるため，対数軸で表している．

精度低下単語は，文書頻度 10を超える部分（赤枠部）では，数が少ないことが見てとれる．最重要単語とその他については，文書頻度が高いところにまで広く分布している．

文書頻度が低い（5以下）箇所に隙間がある理由は，縦軸を対数軸にしており，目盛りが離れているからである．この文書頻度が低い部分に着目するため，Letter value plots[41] に出力し直したものが図 5.2である．縦軸と横軸は図 5.1と同様である．

この図を見ると，精度低下単語の多くは，他の単語と比べて，文書頻度が低い箇所に多く分布していることが見てとれる．

このことから，図 5.2に示すように，文書頻度のしきい値を定め，文書頻度が高い単語を分類に利用することで，精度低下単語の多くを除外することができる．

この文書頻度のしきい値について検討するために，4.1で述べたメール群 1，メール群 2及びメール群 3を用い，文書頻度（df）ごとに単語の種類数の累積比率を求めたのが表 5.1である．例えばしきい値を 4と 5の間にしたとき，最重要単語の約 5割を分類に利用しつつ，精度低下単語の約 8割を除外することができる．

以上より，文書頻度のしきい値を定め，文書頻度の高い辞書なし語を分類に用いることで，精度低下単語の多くを除外しつつ，最重要単語の多くを分類に利用できるため，分類性能の向上が期待できる．次節でこの効果について実験的に確かめる．

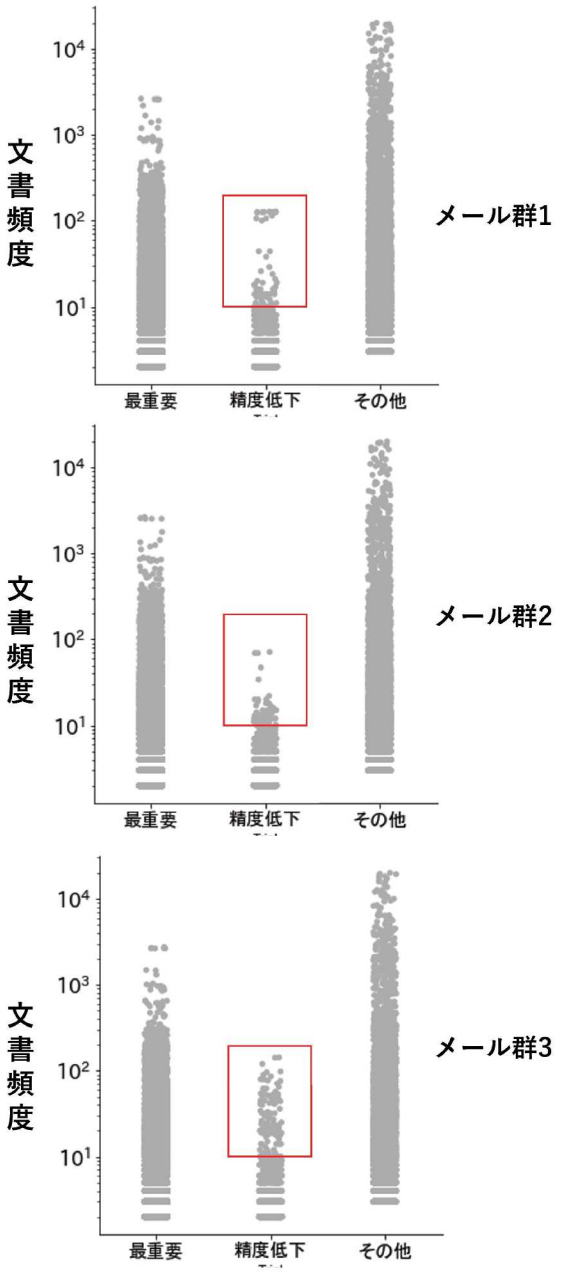

図 5.1: 文書頻度ごとの単語の散布図

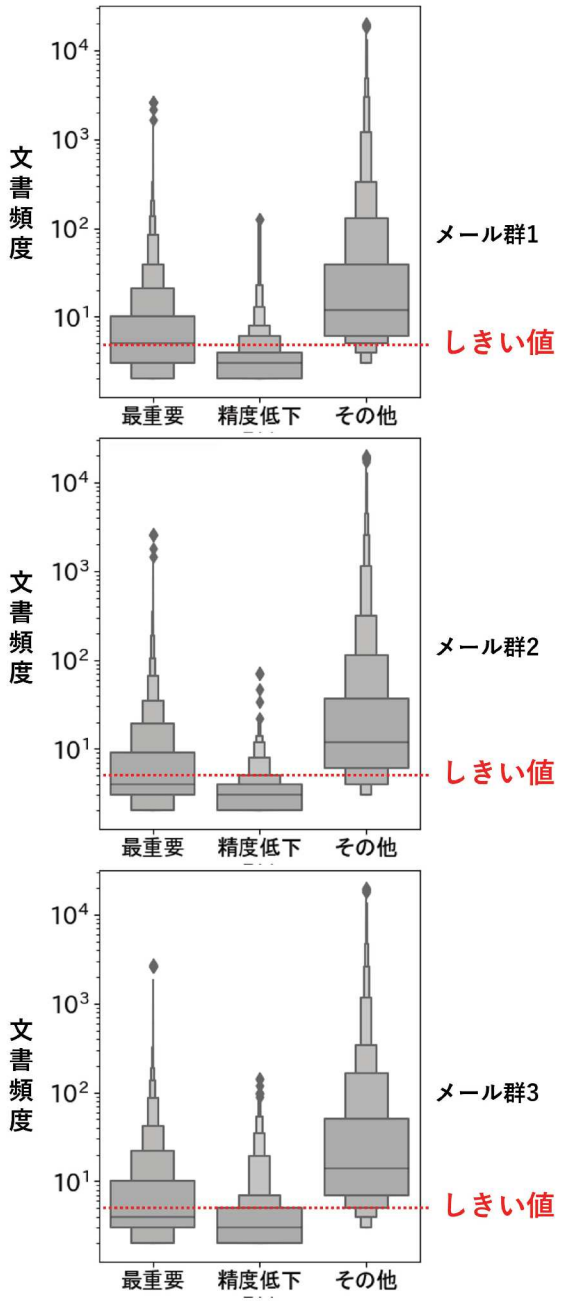

図 5.2: 文書頻度ごとの単語の Letter value plots

表 5.1: 文書頻度ごとの単語の種類数の累積比率（メール群 1）

	メール群 1		
df	最重要	精度低下	その他
1	0.0%	0.0%	0.0%
2	21.3%	44.5%	0.0%
3	37.3%	68.1%	5.1%
4	48.1%	80.5%	11.8%
5	57.8%	86.9%	19.1%
6	63.2%	90.5%	26.1%
7	67.6%	92.5%	31.7%
8	71.5%	94.5%	36.2%
9	74.6%	95.5%	40.5%
	メール群 2		
df	最重要	精度低下	その他
1	0.0%	0.0%	0.0%
2	24.1%	48.3%	0.0%
3	41.4%	70.8%	5.4%
4	52.4%	81.9%	12.8%
5	60.4%	88.3%	19.3%
6	67.2%	91.6%	25.4%
7	71.1%	93.1%	31.2%
8	74.3%	94.8%	35.9%
9	76.5%	95.7%	40.6%
	メール群 3		
df	最重要	精度低下	その他
1	0.0%	0.0%	0.0%
2	23.0%	39.1%	0.0%
3	38.2%	61.4%	5.1%
4	50.6%	74.1%	11.0%
5	58.8%	82.2%	18.3%
6	64.3%	86.0%	24.4%
7	68.7%	88.7%	29.8%
8	71.8%	89.5%	33.6%
9	74.2%	90.2%	37.3%

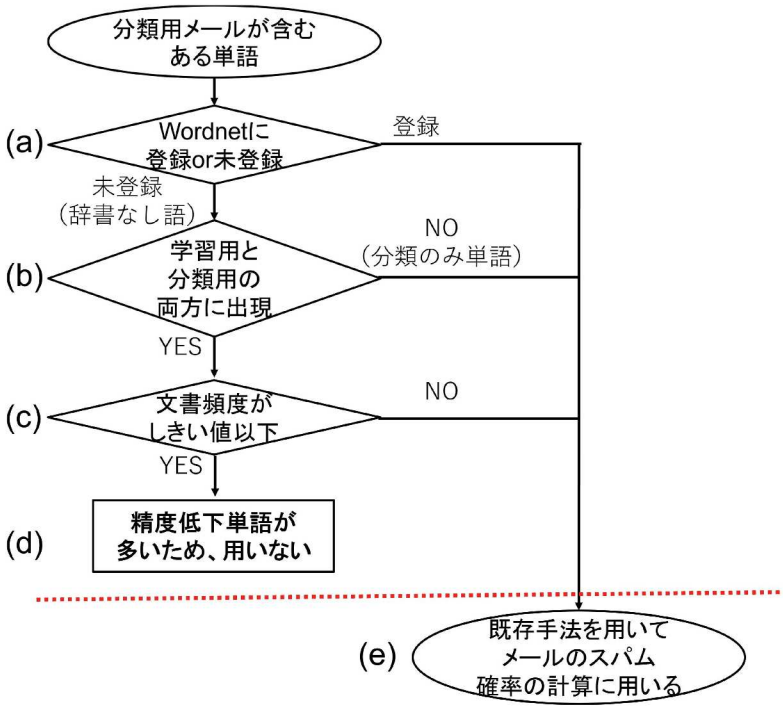

図 5.3: 文書頻度のしきい値による区分を導入した流れ

5.2 辞書なし語の文書頻度による分類性能向上

辞書なし語のうち，図 5.2 で示したしきい値の設定により，精度低下単語の多くを除去したうえで，最重要単語の多くを分類に利用することによって得られる効果を確かめる．

具体的な処理の流れを図 5.3 に示す．すなわち，

(a)： 分類用メールが含む各単語について，Wordnet に登録されているかどうかを確かめ，辞書なし語とそれ以外の単語に分ける．

(b)： 辞書なし語について，学習用と分類用メール群の両方に出現する単語とそうでない単語（分類のみ語）に分ける．

(c)： 学習用と分類用メール群の両方に出現する単語について，

学習用メール群を用いて文書頻度を求め，しきい値以下のものとそれ以外を分ける．

(d)：しきい値以下の語については，精度低下単語を多く含むため，分類に用いないようにする．

(e)：精度低下単語の多くを除外するため，既存手法の分類性能が向上することが期待できる．

文書頻度は，学習用メール群を用い，メール群 1，メール群 2，及びメール群 3 の各々で求め，しきい値については，表 5.1 中の df の 1 から 9 まで変化させて調べた．すなわち，これら 9 通りのしきい値に対して，4.1 で述べたメール群 1，メール群 2，及びメール群 3 を，それぞれ正規とスパムに分けた計 6 つのメール群について，メールのスパム確率の平均値を各々求めた．

適用する既存手法には bsfilter を用い，その実験結果を比較するため，同様の実験をオリジナルの bsfilter でも行った．具体的には，

（しきい値を設定したときの分類結果）−（オリジナルでの分類結果）

の値を求めた．その結果が図 5.4 である．縦軸は，上式で求めたメールのスパム確率の差である．横軸は文書頻度のしきい値である．

結果については，スパムメールは上にあるほど，正規メールは下にあるほどフィルタリング性能が良い．なお，表 5.1 を見ると，文書頻度 1 となる単語は，すべて 0.0%であることから，しきい値 1 の結果はオリジナルのものと一致する．つまり，縦軸の値は 0.0 となる．

図 5.4 を見ると，しきい値が 2 以上のすべての結果について，正規メールの結果（実線）よりも上にスパムメールの結果（点線）があるため，分類性能が向上しているといえる．

しきい値と分類性能の向上具合の関係について詳しく調べるため，図 5.4 の結果について，メール群ごとに

（スパムメールでの分類結果）−（正規メールでの分類結果）

を求めた．この差が正に大きいほど分類性能の向上を示す．その結果が図 5.5 である．縦軸は上式の値を示し，横軸は文書頻度のしきい値である．

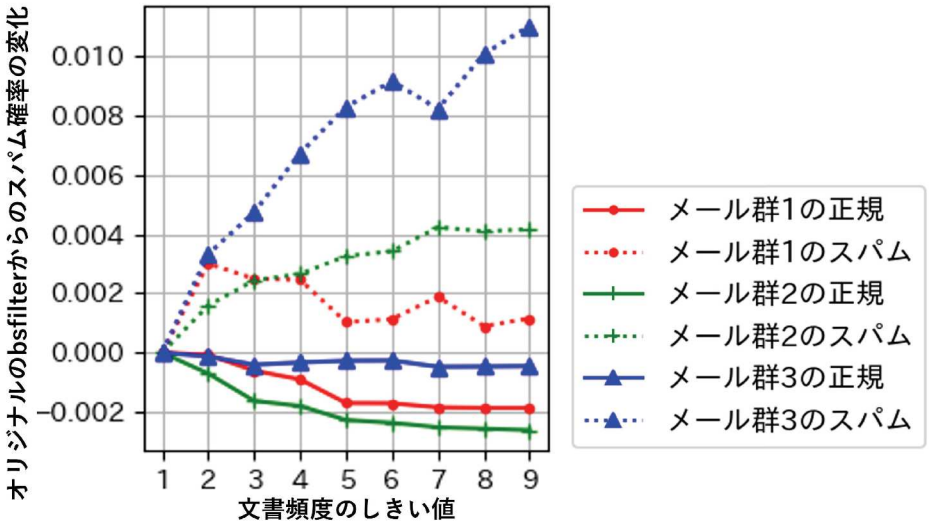

図 5.4: 文書頻度によるしきい値に対するスパム確率の差の平均値

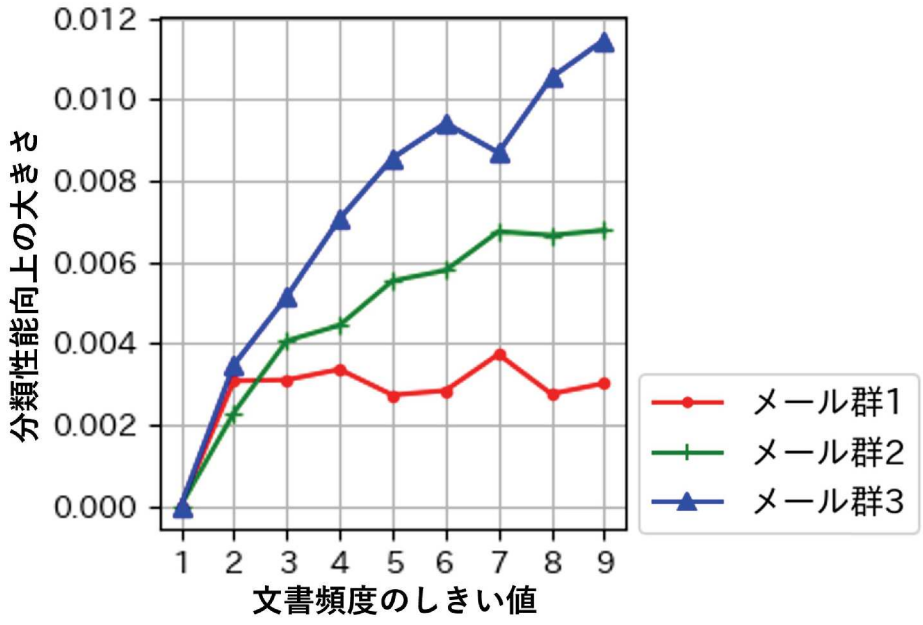

図 5.5: 図 5.4 の結果について，スパムから正規を引いた値

これを見ると，しきい値を 7 付近まで高くしていくと，分類性能の向上が大きくなる傾向にあることが見てとれる．

文書頻度によるしきい値の導入によるスパム確率の変化について，メールごとについても調べるため，メール群 1，メール群 2，及びメール群 3 を用い，正規とスパムメールを分け，文書頻度のしきい値を，例えば 7 としたときのメールの散布図が図 5.6 である．

縦軸はメールのスパム確率の変化量であり，正規メールについては負，スパムメールについて正にあるメールは，分類性能が向上している．

結果を見ると，スパムメールは高い所から低い所にまで広く分布しているため，分類性能が向上したものと低下したものの両方があるということである．正規メールは，スパム確率の変化が± 0.2 以上の箇所に着目すると，0.2 より上（緑枠内）のメール数よりも，-0.2 より下（青枠内）のメール数のほうが多いため，より多くの正規メールを正しく分類できるようになっていると言える．

0.0 付近のメールの数に着目するため，Letter value plots に出力し直した結果が図 5.7 である．

これを見ると，正規メールよりもスパムメールの結果の方が上にある傾向にあることが見てとれる．例えば，中央の 5 割のメールの上部と下部をそれぞれ赤線で繋ぐと，右肩上がりになることが見てとれる．すなわち，正規メールよりも，スパムメールのほうがスパム確率が上昇したメールが多いということであるため，分類性能が向上しているといえる．

以上の結果より，文書頻度にしきい値を定め，文書頻度が高い辞書なし語を分類に用いることで，精度低下単語の多くを除去しつつ，最重要単語の多くを分類に利用することができ，分類性能が向上することを確かめることができた．

比較対象として用いたオリジナルの bsfilter は，分類のみ単語以外の辞書なし語を，未処理のまますべて分類に用いている．この利用手法は，平らの研究 [31] で高性能であることが確かめられている．文書頻度による区分を導入した辞書なし語の利用方法は，これらの利用手法よりも高性能な分類ができている．

この処理は，文書頻度のしきい値に基づく区分を既存手法の利用よりも前（図 5.3 の赤点線よりも上）に行うことから，フィルタリング手法に

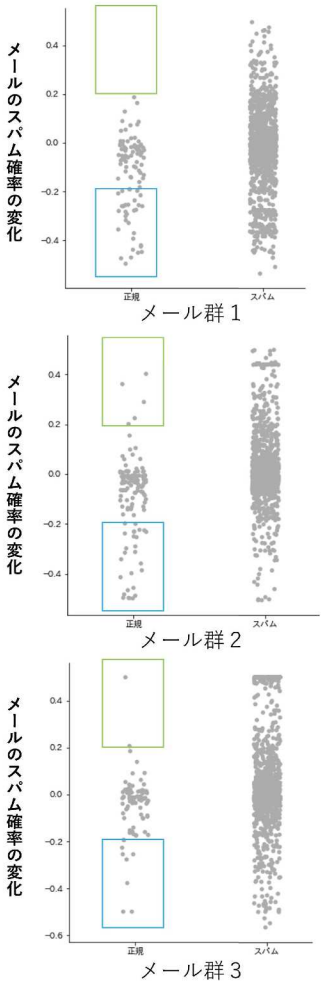

図 5.6: 文書頻度のしきい値を 7 としたときのメールの散布図

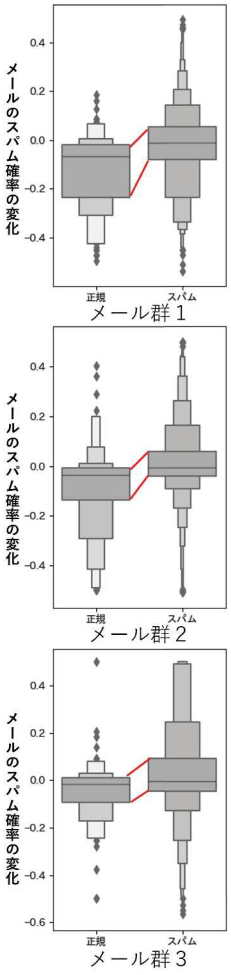

図 5.7: 文書頻度のしきい値を 7 としたときの Letter value plots

依存せずに適用でき，今回の実験結果と同様の性能向上が得られると期待できる．

第6章　分類のみ単語の特性解析と利用手法

この章では，辞書なし語のうち，分類のみ単語に着目した特性解析を行い，これによって見出した特徴を分類に利用する手法を開発する．これを既存手法と併用することで，分類性能のさらなる向上を試みる．

6.1　データセットの区分，及び利用方法

一般にメールの特徴は時間の経過とともに変化する．正規メールは，活動状況，話題，及び交友関係等に関わる本文の変化によっても特徴が変化する．スパムメールについては，トレンドの変化に加え，3.2.1で述べた辞書なし語の作成や置き換えによっても特徴が変化する．

この特徴の変化は，学習用メール群から時間が経過した分類対象のメールほど起こりやすく，学習用メール群（過去に受信したメール群）にない分類のみ単語も増加しやすいと考えられる．

このような学習用メール群からの時間経過について着目した解析を行うため，データセットの区分を，図6.1に示すように，正規とスパムメールのそれぞれについて，受信日時の古い2割のメール群を学習用，以後のものを分類用メール群として，受信日時の昇順に並べたのちに8等分し，1から8の番号を割り当てた．メール群を割合によって区分することで，3.1で述べたSpamAssassinとTRECのように，構成するメール数が異なるメール群の結果が比較しやすくなる．

実際にフィルタを運用する際には，一般に新しい受信メールも追加して機械学習を更新するが，本研究では，受信メールの時間経過による特徴の変化に着目するため，学習用メールを固定して解析に用いた．

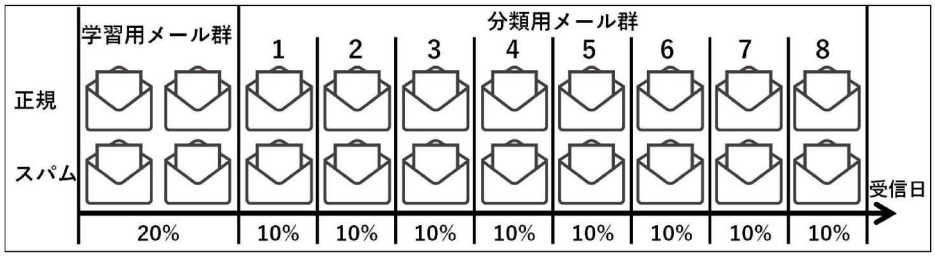

図 6.1: 学習用メール群と分類用メール群の区分

6.2 正規メールとスパムメールの分類性能の時間的変化

学習用メール群からの時間経過によるメールの特徴の変化が，分類性能に与える影響について調べるため，3.1で述べたSpamAssassinとTRECについて，分類用メール群の1から8をそれぞれ用い，bsfilterによる分類実験を行った．その結果を図6.2に示す．

この図の横軸は，分類用メール群の番号を示しており，番号が大きい方が，学習用メール群の受信日からの経過時間が長い．縦軸は，分類用メール群ごとに求めたメールのスパム確率の平均値であり，正規，及びスパムメールのそれぞれについて示している．

これを見ると，SpamAssassin，及びTRECは同じ傾向を示しており，正規メールについては，スパム確率が0.0付近と低いままであることから，時間経過に依存せず，分類性能が高いことが見てとれる．

一方で，スパムメールについては，時間の経過が最も短い分類用メール群1でも0.85付近と低く，その後0.6付近にまで分類性能が低下していくことが見てとれる．すなわち，正規よりもスパムメールのほうがメールの特徴が変化しやすいということである．

この原因の一つとして，3.2.1で述べたスパム送信者による，辞書なし語の作成行為が挙げられ，これにより，辞書なし語の分類のみ語も増加傾向にあることが考えられる．

このことを確かめるため，次節の解析を行った．

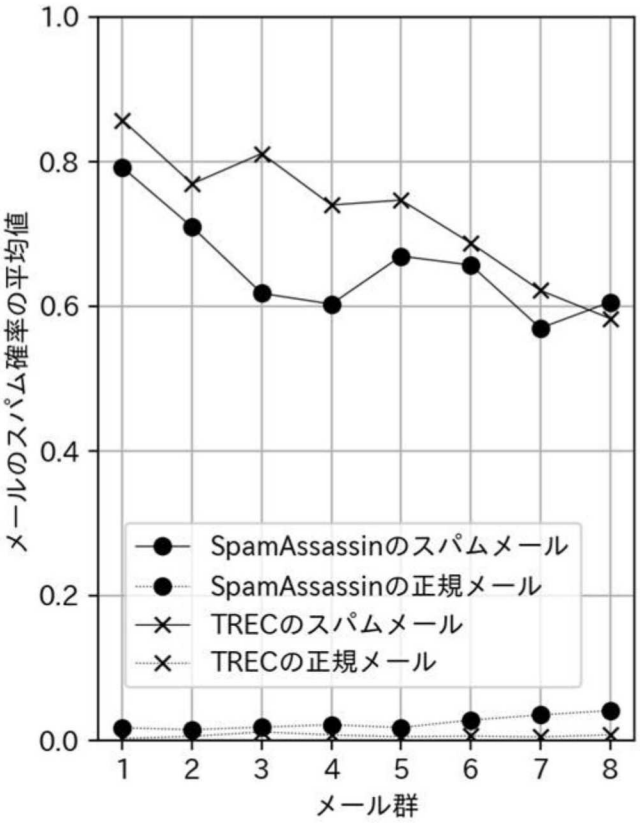

図 6.2: 日数の経過に対する分類性能の変化

6.3 分類のみ語の活用の可能性

辞書なし語の分類のみ単語の数について調べるため，2.1で示したSpamAssassinとTRECの両方のデータセットを用い，図6.3に示すように，分類用メール群が含む単語のうち，学習用メール群にない単語（分類のみ単語）とそうでない単語（学習用メール群が含むため，学習済となる）に分け，その各々で種類数を集計した．その結果を表6.1に示しており，SpamAssassinとTRECは共通して，学習済よりも，分類のみの単語がかなり多いことがわかる．

この分類のみ単語のメール分類への活用可能性を調べるため，単語の

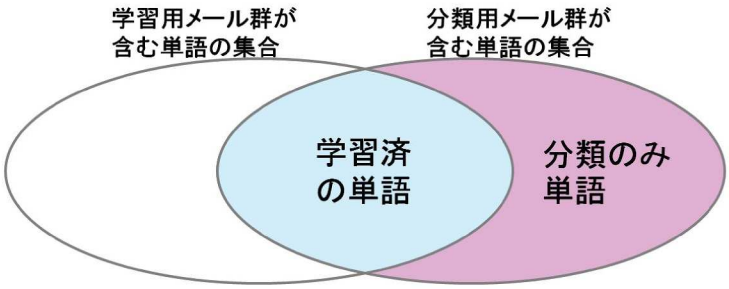

図 6.3: 分類のみ単語と学習済単語の関係

表 6.1: 学習済と分類のみ単語の種類数

	SpamAssassin	TREC
学習済	26,049	102,558
分類のみ	81,087	627,751

種類数の時間的変化を調べてみた．その結果を図 6.4 に示す．

横軸は，分類用メール群の番号であり，番号が大きい方が，学習用メール群の受信日からの経過時間が長い．縦軸は，1 通のメールに出現する単語の種類数のうち，分類のみ語の種類数が占める割合を求め，分類用メール群ごとに平均した値である．分類のみ語を辞書なし語と辞書あり語に分け，正規とスパムのそれぞれで割合を求めている．

この図を見ると，SpamAssasssin と TREC はともに，メールの件名や本文が含む分類のみ語について，スパムメールの辞書なし語が多く，かつ増加傾向にあることが見てとれる．

すなわち，分類対象のメールに辞書なし語の分類のみ単語が多いとき，そのメールがスパムメールである傾向が強く，それは日数の経過に従って強くなるといえる．

この傾向をスパムメールの特徴として分類に利用するため，次節の実験を行った．

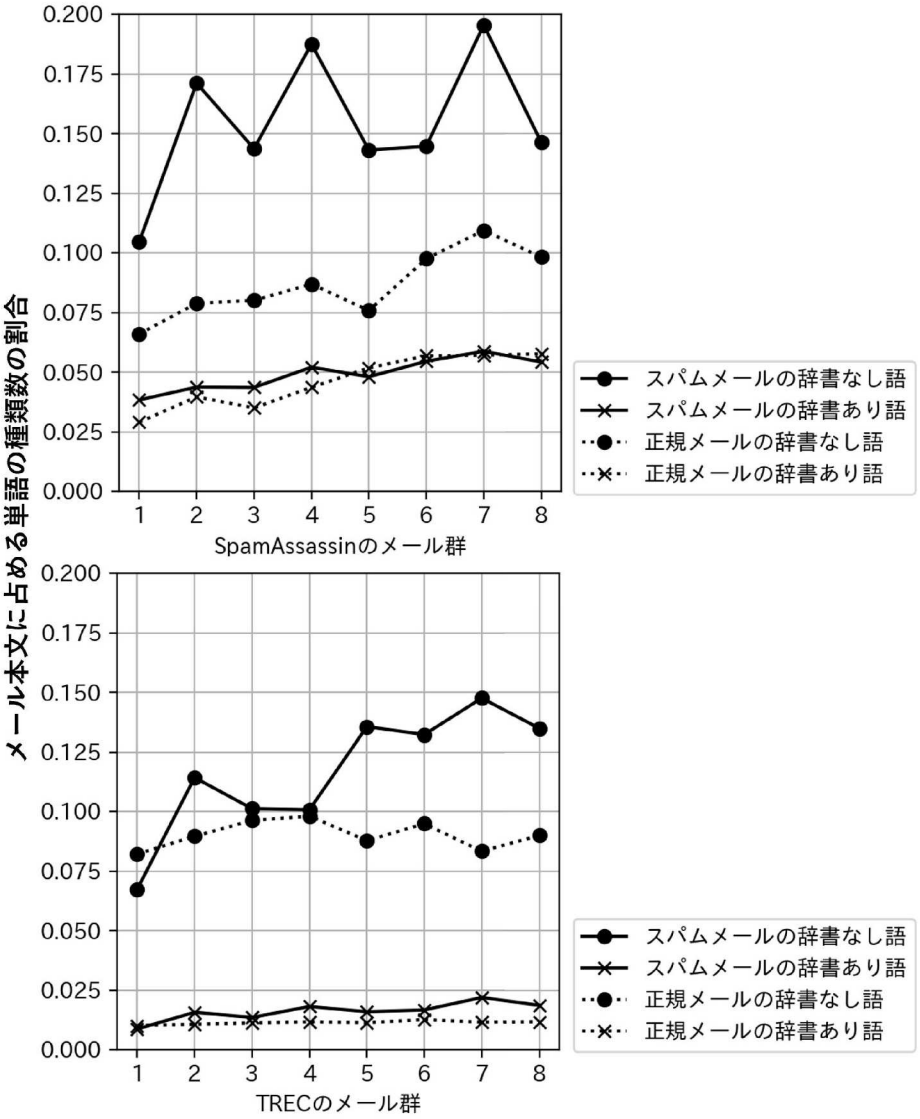

図 6.4: メール 1 通あたりに占める分類のみ語の種類数の割合の時間的変化

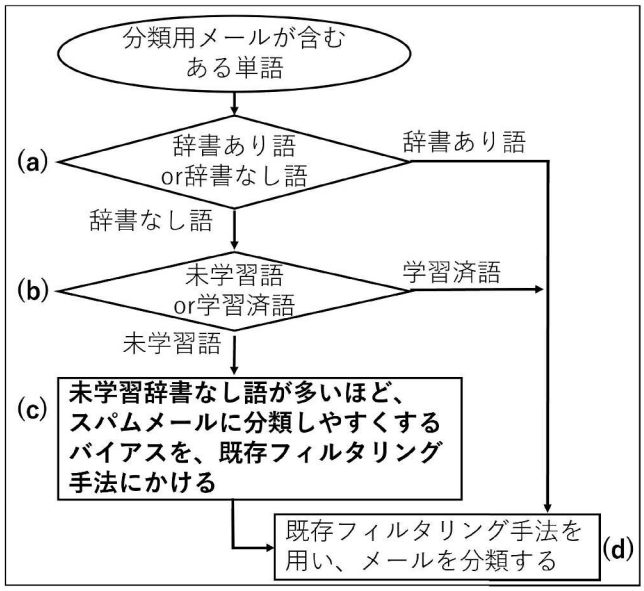

図 6.5: 既存フィルタリング手法への辞書なし語の分類のみ単語の分類適用

6.4 辞書なし語の分類のみ単語の利用

6.4.1 既存フィルタへの利用

辞書なし語の分類のみ単語を既存手法に適用するには，分類用メールに辞書なし語の分類のみ単語が多く出現するほどスパムメールに分類し易くなるようにバイアスをかければよい．

具体的には，図 6.5 に示す流れで処理を行う．すなわち，

(a)： 分類用メールが含む各単語について，辞書なし語と辞書あり語に分ける．

(b)： 辞書なし語について，分類のみ語（学習用メール群が含まない単語）とそれ以外（学習用メール群が含む単語）を分ける．

(c)：辞書なし語の分類のみ語を集計し，その数が多いほど，スパムに分類し易くするバイアス処理を，既存フィルタリング手法に導入する．

(d)：辞書なし語の分類のみ語以外については，既存フィルタリング手法を用いてメール分類に利用する．

この処理は，各々のフィルタリング手法に合わせて適用できる．その具体例として，2.2.5で述べたSVM[8]や2.2.6で述べたBONSAI[9]を用いたフィルタリング手法へ適用する手法について説明する．

メールフィルタリングで用いられているサポートベクトルマシン（SVM）は，図6.6の左に示すように，多次元の単語の特徴ベクトル空間上に付置したメール群を正規とスパムに二分するため，最も近くにある正規とスパムメールの距離（マージン）を最大化する分離超平面を求める．

このSVMに辞書なし語の分類のみ単語の利用を適用するためには，図6.5の(c)の処理について，図6.6の右に示すように，分類用メールごとに辞書なし語の分類のみ単語の量を求め，その量が多いほど，分類用メール位置を各々，スパムメールの方向に移動させるようにバイアスをかける処理を行えばよい．

決定木に基づくBONSAIを用いたメールフィンタリング手法は，図6.7の左に示すように，学習用メール群が含む全ての単語について，スパムに出現する確率を求め（スパムほど高い），メールを変換する．さらに，この確率を，2.2.6で述べた表と照らし合わせ，「a，b，k，x，y」の5種類のアルファベットに変換する．BONSAIは，この並びについて，正規またはスパムに特徴的なパターンを学習する．

このBONSAIを用いたメールフィルタリング手法に，辞書なし語の分類のみ語の利用を適用するためには，図6.5の(c)の処理について，図6.7の右に示すように，辞書なし語の分類のみ単語に対し，ある程度高い確率を設定するようなバイアスをかける処理を行えばよい．

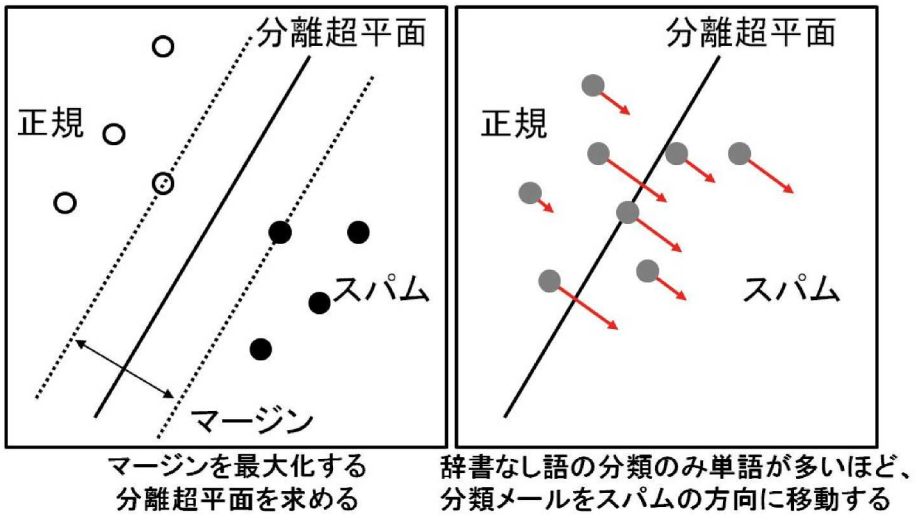

図 6.6: SVM への辞書なし語の分類のみ単語の分類適用

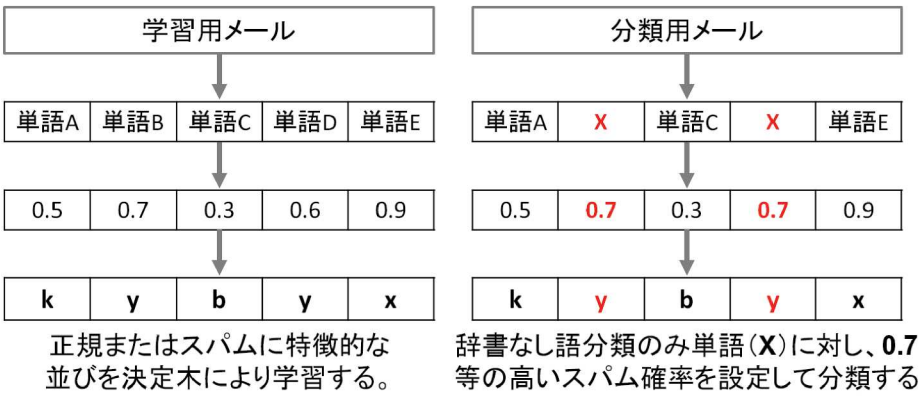

図 6.7: BONSAI への辞書なし語の分類のみ単語の分類適用

本研究では，現在広く用いられている bsfilter に辞書なし語の分類のみ単語を適用し，実験的にその効果を調べる．具体的には，図 6.5 の (c) の処理を，図 6.8 に示すように，辞書なし語の分類のみ単語に対してスパム確率を一律に設定するようにする．すなわち，

(a)： 分類用メールが含む各単語について，Wordnet に登録されているかどうかを確かめ，辞書なし語と辞書あり語に分ける．

(b)： 辞書なし語について，分類のみ語（学習用メール群が含まない単語）と学習済語（学習用メール群が含む単語）を分ける．

(c)： 辞書なし語の分類のみ単語について，bsfilter の分類のみ語に対する処理を行わず，バイアスとなるスパム確率を決め，一律に設定する．

(d)： 辞書なし語の分類のみ単語以外の単語については，bsfilter の処理により，スパム確率を求める．

これを改変後 bsfilter と呼ぶ．(c) の処理は，2.2.2 で述べた Gary Robinson 方式の式（2.2）で用いられている，分類対象に初めて出現する単語のスパム確率 x を，辞書なし語の分類のみ語について一律に設定することに相当する．

ここで設定する最適なスパム確率を探索するため，辞書なし語の分類のみ単語にスパム確率を一律に設定し，これを変化させながら分類実験を行い，この結果をオリジナルの bsfilter の結果と比較する．

bsfilter では，辞書なし語の分類のみ単語に対し，2.2.4 で述べた，1.〜5. の記号等の削除や，大文字小文字の変換による処理を行うことで，学習済みデータベースから単語を照合し，一致したものを当該分類のみ単語に当てはめて分類に利用する．すなわち，辞書なし語の分類のみ単語の一部しか扱っていない．また，辞書あり語と辞書なし語の区別も行っていない．

6.1 で示したデータセットを用い，辞書なし語の分類のみ単語のスパム確率を，1.0（最大値）に設定したときの分類結果は図 6.9 のようになる．縦軸はメールのスパム確率の平均値であり，比較のためにオリジナルの bsfilter の結果も示している．横軸はメール群の番号である．

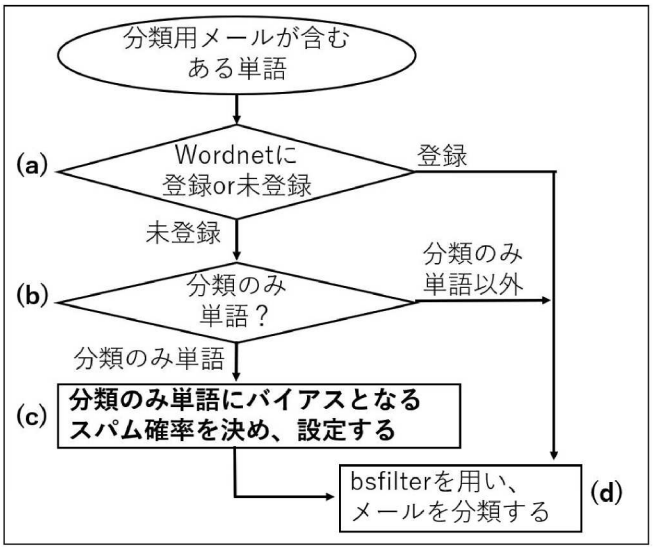

図 6.8: 改変後 bsfilter の処理の流れ図

辞書なし語の分類のみ単語のスパム確率を 1.0 に設定し，分類に利用したことにより，メールのスパム確率はオリジナルの bsfilter よりも高くなる．実際に図 6.9 では，SpamAssassin と TREC はともに，オリジナルの bsfilter の結果よりも，正規とスパムメールの両方でスパム確率が高くなっている．

このメールのスパム確率の上昇は，スパムよりも正規メールのほうが大きい．スパムメールのスパム確率の上昇は改善を示し，正規メールのスパム確率の上昇は悪化を示すため，この場合は，分類用メール全体での分類性能は悪化する．

これと同様の実験を，辞書なし語の分類のみ単語に設定するスパム確率を 0.0 から 1.0 まで変化させながら行うことで，分類性能が最も向上する最適な値の探索を試みた．その結果を図 6.10 に示す．

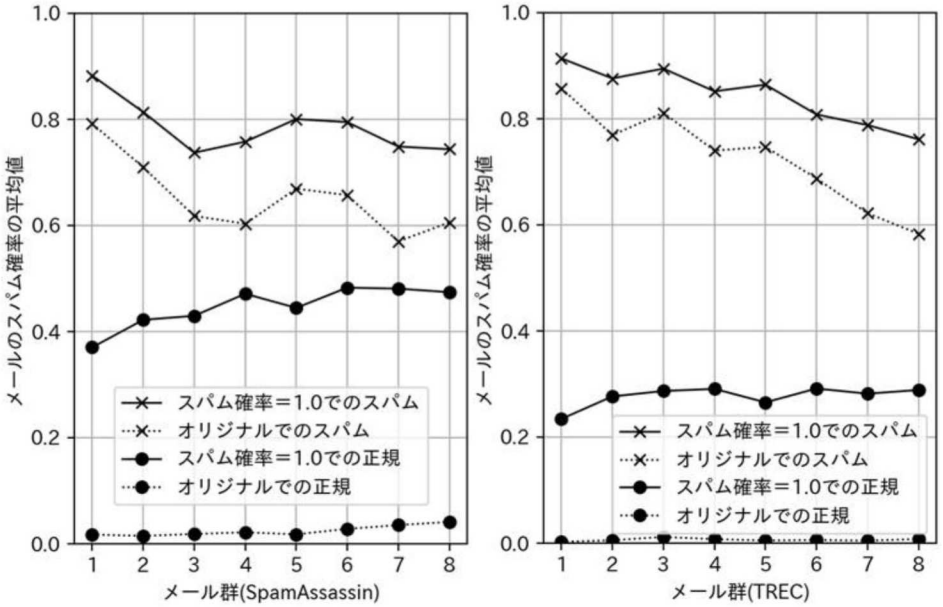

図 6.9: 改変後 bsfilter の辞書なし語の分類のみ単語のスパム確率を 1.0 としたときとオリジナル bsfilter との分類性能比較

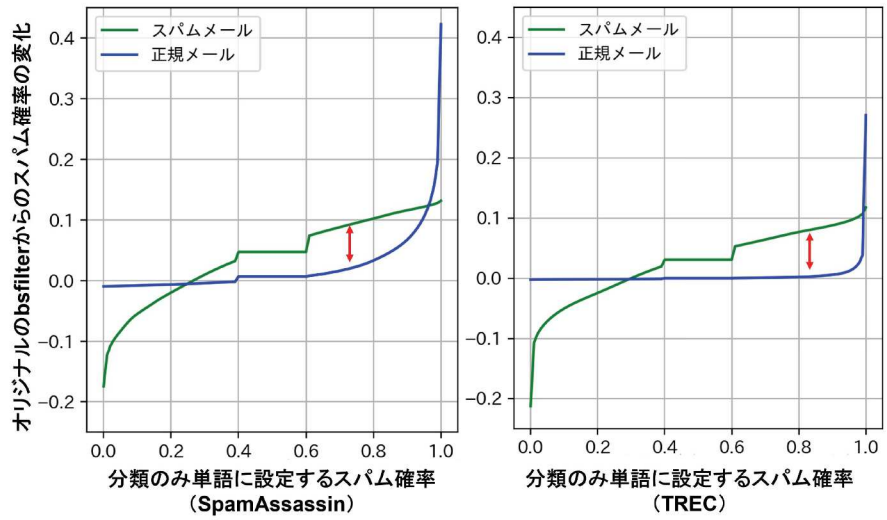

図 6.10: 辞書なし語の分類のみ単語のスパム確率と分類性能の関係

横軸は，設定したスパム確率であり，縦軸は，メールのスパム確率について，

(改変後のスパム確率) – (オリジナルのスパム確率)

を分類用メール群 1から 8で求め，これらを平均した値を示している．上式は，メールのスパム確率が高くなっていれば正，低くなっていれば負となる．すなわち，スパムメールのスパム確率が，正規メールのスパム確率と比べて，上にあるほど（図 6.10 の赤矢印が離れているほど）良い分類結果が得られるということである．

最も良い分類結果が得られるスパム確率を探索するため，設定するスパム確率ごとに，

(スパムメールのスパム確率) – (正規メールのスパム確率) × 2

を調べ，図 6.10 と同様の形式で表したものが図 6.11 である。ここで，正規メールのスパム確率を 2 倍している理由は，正規メールのスパム確率の上昇が小さくなるスパム確率を，探索しやすくするためである．メールフィルタリングでは，正規メールを多く受信することが最優先であるが，上記の操作をすることで，正規メールの分類性能を高く保ちつつ，スパムメールの分類性能を向上できるスパム確率を探索できる．

この図を見ると，SpamAssassin では，スパム確率が 0.63 の時に最大値（0.06）となり，TREC では，スパム確率が 0.90 の時に最大値（0.08）となることが見てとれる．本研究では，一般性のあるスパム確率を探索するため，これら 2 つの結果を合わせ，各スパム確率ごとに平均値を求めた．その結果が図 6.12 である．これを見ると，スパム確率が 0.71 の時に最大値（0.06）となることが見てとれる．スパム確率 0.7 の周辺で，結果がほとんど同じであることから，良い分類結果を得るにはスパム確率を 0.7 付近に設定するとよい．

スパム確率 0.7 のときの分類性能の時間的変化を，図 6.9 と同じ形式で表すと，図 6.13 のようになる．オリジナルからのメールのスパム確率の変化について，SpamAssassin と TREC の結果はともに，正規メールのスパム確率の上昇よりも，スパムメールのスパム確率の上昇のほうが大きいため，分類用メール群の受信日によらず，分類性能が向上している．

分類性能についてさらなる検証を行うため，bsfilter を用い，データセットの区分を変えながら分類精度を確かめ，その結果について次節で報告する．

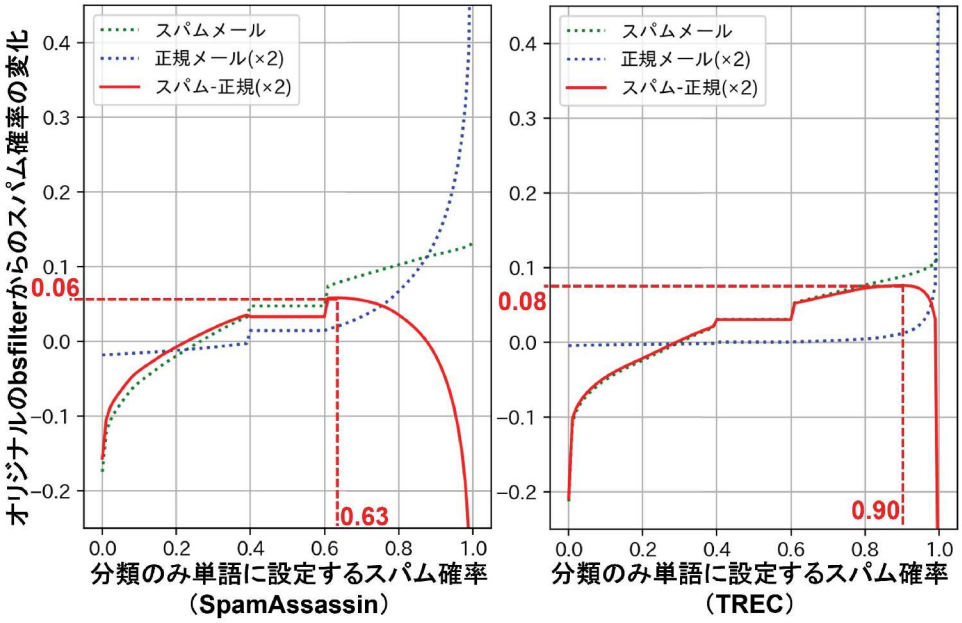

図 6.11: スパムメールと正規メールのスパム確率の差

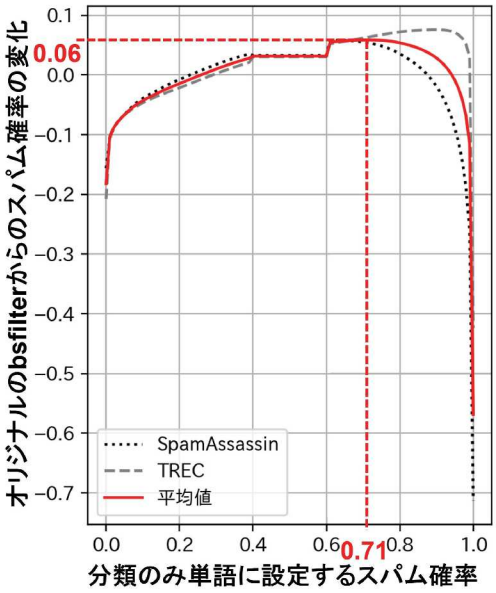

図 6.12: 分類のみ単語に設定するスパム確率と分類性能の変化の関係

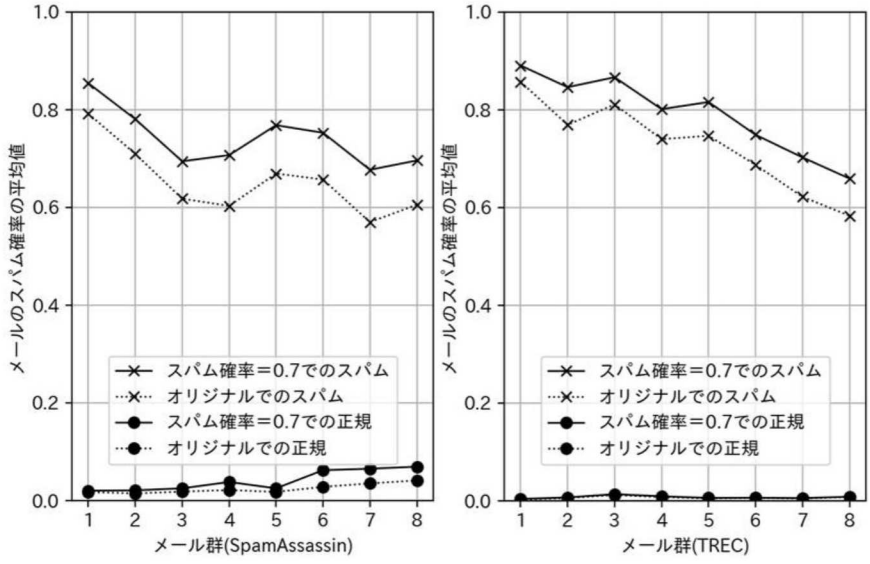

図 6.13: 改変後 bsfilter の辞書なし語の分類のみ単語のスパム確率を 0.7 としたときとオリジナル bsfilter との分類性能比較

6.4.2 辞書なし語の分類のみ単語利用による分類性能の変化

分類用メール群の受信日の経過と，分類性能の変化の関係を調べるため，図 6.14 に示すように，最も古い 2 割のメール群を学習用とし，以後のものを 4 等分して分類用とし，1 から 4 の番号を割り当てた．

分類精度の評価には，3.3 で述べた ROC 曲線を用いた．例えば，TREC のメール群で，分類用メール群 1 から 4 をすべて分類したときの ROC 曲線は図 6.15 のようになる．横軸は偽陽性率を示しており，縦軸は真陽性率である．オリジナルよりも，スパム確率を 0.7 とした改変後 bsfilter のほうが AUC の値が大きいため，分類精度が高くなっていることがわかる．

これと同様の実験を，TREC と SpamAssassin を用いて行い，学習用メール群で学習し，分類用メール群 1 から 4 の各々について，ROC 曲線が描く AUC の値を求めた．その結果を図 6.16 に示す．横軸はメール群の番号であり，番号が大きいほど分類用メール群の受信日が新しい．縦軸は AUC である．

この実験を，2.2.4 で述べた bsfilter の（1）～（5）の記号等の削除や，大文字小文字の変換の処理を行わないように改変した bsfilter（未使用）でも行った．この結果も比較対象として追加することで，スパム確率を 0.7 とした改変後 bsfilter と，オリジナルの bsfilter の辞書なし語の分類のみ単語の利用効果をそれぞれ調べることができる．

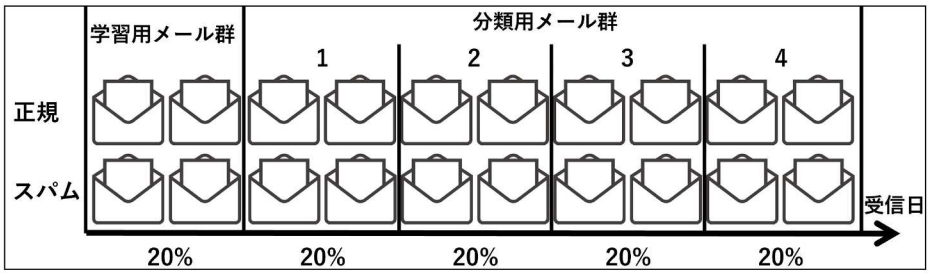

図 6.14: 学習用メール群と分類用メール群 1 から 4 の区分

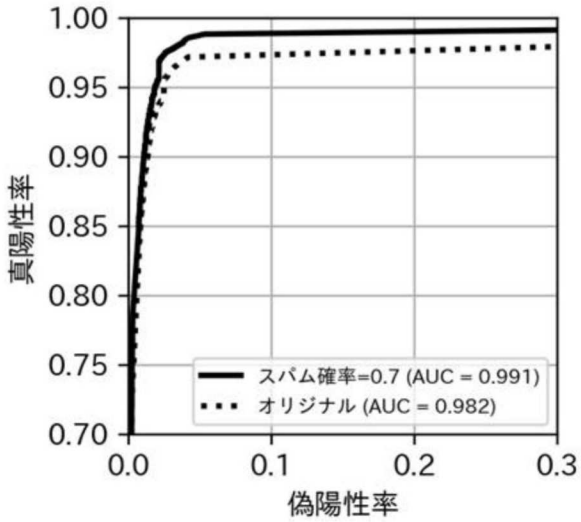

図 6.15: TREC のパターン 1 で求めた ROC 曲線と AUC（真陽性率が 0.7 未満または偽陽性率が 0.3 を超える部分は図の拡大のため省略）

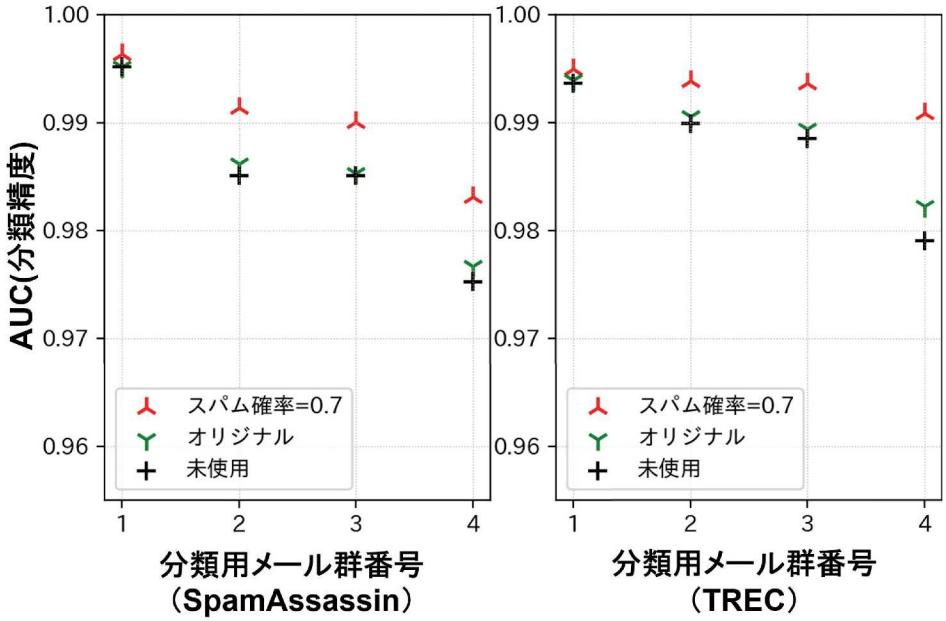

図 6.16: 分類用メール群の受信日と分類精度の関係

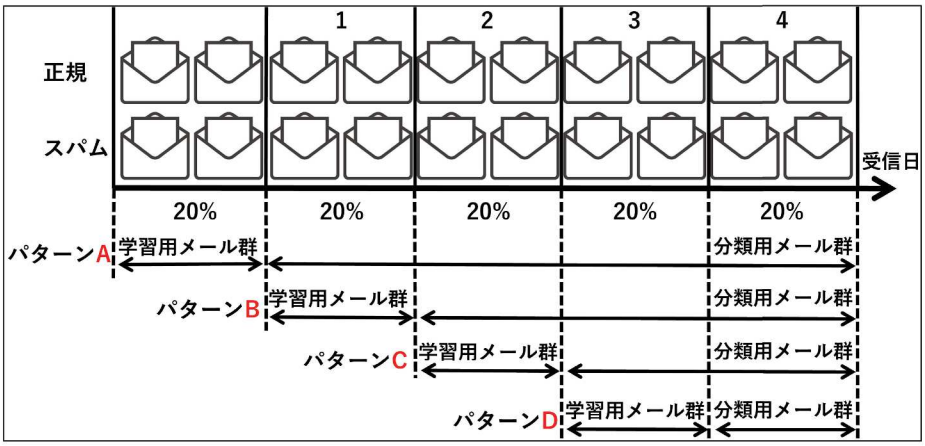

図 6.17: 学習用メール群を変えた A から D の区分パターン

この図を見ると，SpamAssassin と TREC の両方の結果において，右に行くほど（分類用メール群の受信日が新しいほど）分類精度が低下する傾向にあるが，スパム確率を 0.7 に設定すると，他の結果と比べて，分類精度が下がりづらく，かつ最も高精度であることが見てとれる．

この精度向上の効果は，分類のみ単語に 0.7 という中間（0.5）よりも高いスパム確率を設定，すなわちスパムの特徴として利用したことで得られている．このことから，図 6.2 で確かめたスパムメールの分類性能の低下を抑えていることが考えられ，スパム送信者が辞書なし語を新たに作成する行為等の対策として機能することが期待できる．

異なる学習用メール群を用いてみるため，図 6.17 に示すように，これまでの解析時に用いたパターン A の他に，学習用メール群をメール群 1 としたパターン B，メール群 2 としたパターン C 及びメール群 3 としたパターン D を用意した．分類用メール群は，学習用メール群よりも新しいメールである．

図 6.16 と同様の形式で，TREC と SpamAssassin を用い，パターン A から D の各々で分類実験を行い，ROC 曲線が描く AUC の値を求めたものが図 6.18 である．

横軸はメール群のパターン記号であり，A から D に進むほど，学習用メール群の受信日が新しい．実験結果を比較するため，6.16 で用いた

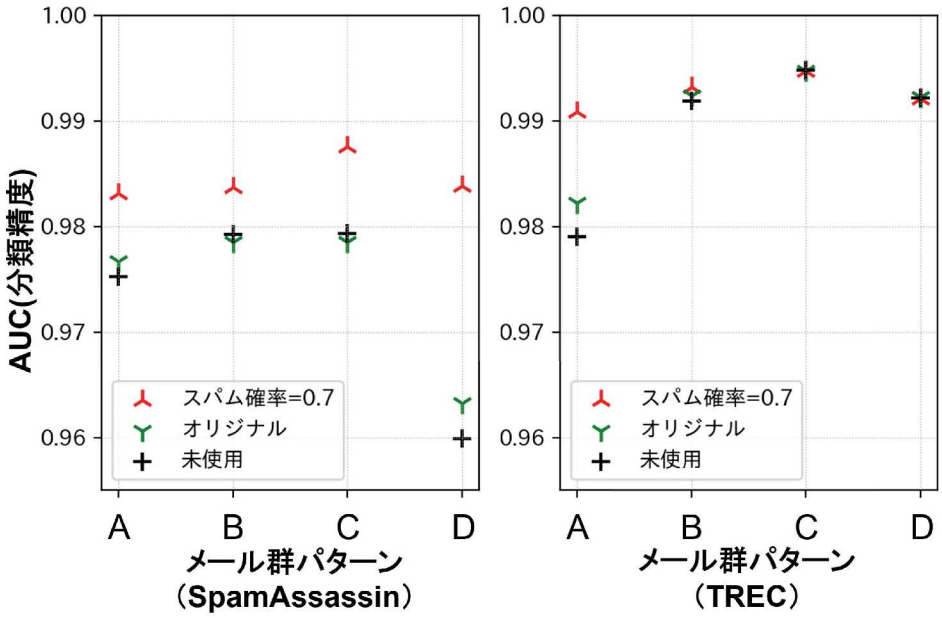

図 6.18: 学習用メールと分類精度の関係

bsfilter（未使用）も用いて分類実験を行った．

この図を見ると，スパム確率を 0.7 に設定したときの分類精度は，SpamAssassin ですべて 0.98 以上，TREC で 0.99 以上と高く，学習用メール群の受信日によらず高精度であることが見てとれる．

TREC の結果では，学習用メール群の受信日を新しくしていくと，すべての手法の分類精度が重なることが見てとれる．この原因は，TREC の受信期間が約 3 か月と短く，学習用と分類用メール群の受信日が近いため，分類用メールに新しい特徴が出現しづらく，分類のみ語も少なくなったことが考えられる．

以上の結果より，分類のみ単語に施す処理は，オリジナルの bsfilter が行っている，2.2.4 で述べた（1）～（5）の記号等の削除や大文字小文字の変換よりも，本論文で提案するスパム確率を 0.7 に設定する方法が高精度であるといえる．

分類性能のさらなる検証のため，辞書なし語の分類のみ単語を使用し

ないように改変した bsfilter （図 6.18 の実験で用いたものと同様）を比較対象として，辞書なし語の分類のみ単語に，スパム確率 0.7 を設定した場合のスパム確率の変化をメールごとに調べた．

パターン 1 のデータセットを用い，SpamAssassin と TREC の両方について，スパム確率の変化について分類用メールの分布を調べた．この結果を散布図（図 6.19）と Letter value plots（図 6.20）で示す．

縦軸は，辞書なし語の分類のみ単語の利用について，

（スパム確率に 0.7 を設定した場合）－（利用しない場合）

でのメールのスパム確率の変化を求めた値である．

図 6.19 を見ると，ほぼすべてのメールでスパム確率が上昇していることが見てとれる．この理由は，改変後 bsfilter で，辞書なし語の分類のみ単語をスパムの特徴として利用しているからである．さらに，図 6.20 を見ると，SpamAssassin と TREC の両方について，正規メールは，0.0（スパム確率が変化していない）にメールが多く（SpamAssassin で約 55%，TREC で約 95%），その他のメールについても，スパム確率の変化が小さいことが見てとれる．他方，スパムメールは，正規メールよりもスパム確率の上昇が多く得られていることから，これが今回の実験における分類精度向上（図 6.18）の原因と考えられる．

スパム確率が大きく上昇した SpamAssassin のメールについて，正規メール（図 6.19 の（a））は，図 6.21 に示すような URL や日付を含む短いニュースメールを含んでおり，スパムメール（図 6.19 の（b））は，図 6.22 に示すような URL や記号といった辞書なし語の分類のみ単語を用いたメールを含んでいる．

TREC について，正規メール（図 6.19 の（c））は，図 6.23 に示すような英語以外のメールを含んでおり，スパムメール（図 6.19 の（d））は，図 6.24 に示すような HTML タグを用いた広告メールを含んでいる．

このことから，辞書なし語の分類のみ単語にスパム確率 0.7 を設定することで，スパム確率が上昇する正規メールもあるが，これよりも多くのスパムメールを正しく分類できるため，分類性能の向上を得ることができている．

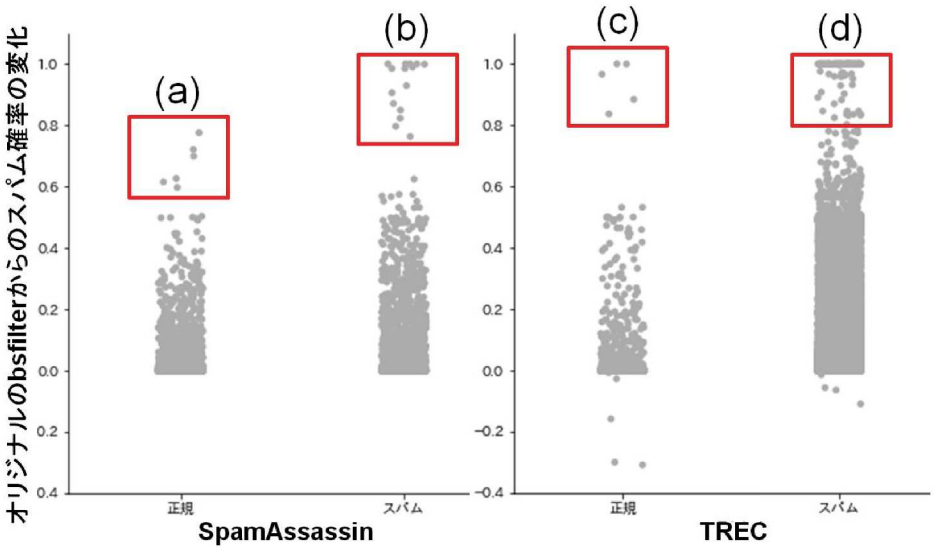

図 6.19: 辞書なし語の分類のみ単語利用によるメールのスパム確率の変化（散布図）

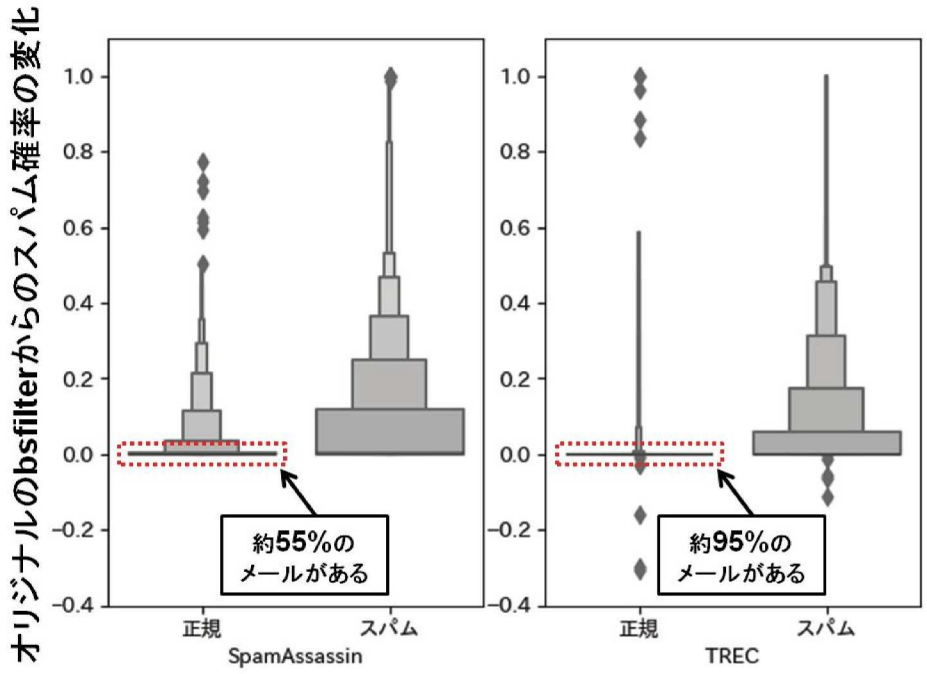

図 6.20: 辞書なし語の分類のみ単語利用によるメールのスパム確率の変化（Letter value plots）

```
Tibetans see hint of detente with China
URL: http://www.newsisfree.com/click/-2,8418827,215/
Date: 2002-10-01T04:33:57+01:00
*World latest:* Dalai Lama's envoy hails first contact in 20 years.
```

図 6.21: スパム確率が大きく上昇した正規メールの例 (Spamassassin)

```
* * * * * FREE Personals * * * * *
For a limited time, we are offering all our customers
the option to place their own adults only personals
(PHOTO OR NON-PHOTO) ad at LAXPress.com AT NO CHARGE!
* * * * * FREE LIVE NUDE CHATROOMS | NO CHARGE FOR ACCESS ! * * * * *
http://www.laxpress.com/50SITES/adultupskirts.html
http://www.laxpress.com/50SITES/allpetite.html
http://www.laxpress.com/50SITES/amateurfreedom.html
http://www.laxpress.com/50SITES/amateuruniversity.html
http://www.laxpress.com/50SITES/analdebutants.html
http://www.laxpress.com/50SITES/asianexxxstacy.html
http://www.laxpress.com/50SITES/teenmaidens.html
http://www.laxpress.com/50SITES/totallytits.html
http://www.laxpress.com/50SITES/uniformboys.html
http://www.laxpress.com/50SITES/voyeuraddicts.html
http://www.laxpress.com/50SITES/voyeurdorm.html
http://www.laxpress.com/50SITES/voyeurlounge.html
```

図 6.22: スパム確率が大きく上昇したスパムメールの例 (Spamassassin)

```
BUTUH UANG TUNAI YG CEPAT & MUDAH...???
Wujudkan kebutuhan Anda utk pendidikan putra-putri, renovasi rumah, modal
usaha, menikah atau melunasi kartu kredit.Untuk itu,kami dapat membantu
Anda melalui Kredit Tanpa AgunanKU (PersonalLoans) dari HSBC. Proses 7 s/d 14 hr krj.
Syarat penghasilan kotor 3,5 jt/bln cukup dgn melampirkan Slip Gaji + KTP
atau pemilik kartu kredit apapun min limit 5 jt (masa keanggotaan 6 bln)
```

図 6.23: スパム確率が大きく上昇した正規メールの例 (TREC)

```
download Photoshop CS3 right today for only $89!
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<HTML><HEAD>
<META http-equiv=Content-Type content="text/html; charset=iso-8859-1">
<META content="MSHTML 6.00.2900.2180" name=GENERATOR>
<STYLE></STYLE>
</HEAD>
<BODY bgColor=#ffffff>
<DIV><FONT face=Arial size=2><IMG alt="" hspace=0
src="cid:001101c790f6$72de4f78$_CDOSYS2.0" align=baseline
border=0></FONT></DIV>
```

図 6.24: スパム確率が大きく上昇したスパムメールの例 (TREC)

第7章　おわりに

本論文では，スパム送信者がフィルタすり抜けのために，辞書なし語を悪意を持って作成し続けていることに着目し，辞書なし語の特性解析を行い，これによって見出した特性を分類に利用する手法を開発した．

開発手法の全体像を図 7.1 に示す．分類用メールが含む単語の辞書なし語のうち，分類のみ単語とそれ以外に分ける必要性を 4 章で示し，分類のみ単語以外に対する処理を 5 章で提案し，分類のみ単語に対する処理を 6 章で提案した．

4 章では，辞書なし語が分類性能に与える影響について調べるため，辞書なし語，名詞，動詞及び形容詞の各々について，bsfilter を用いた分類実験を行った．これにより，辞書なし語の分類性能は，他の品詞と比べて分類性能に与える影響が大きいことを確かめた．

この原因について調べるため，辞書なし語を 3.5 で述べた最重要単語，精度低下単語，学習のみ単語，分類のみ単語及びその他の出現パターンに分け，その各々について単語の種類数を集計して比較した．これにより，辞書なし語は，学習用と分類用メール群の両方に出現する単語（最重要単語，精度低下単語及びその他）のうち，最重要単語が多いことを確かめた．また，辞書なし語は他の品詞と比較して，分類のみ単語と学習のみ単語がかなり多いことを確かめた．

以上の結果より，辞書なし語について，分類のみ単語とそれ以外の単語に区分して扱う（図 7.1 の赤枠部）ことによって分類性能が向上する可能性を見出した．

5 章では，辞書なし語のうち，学習用と分類用メール群の両方に出現する単語，すなわち最重要単語，精度低下単語及びその他について，精度低下単語の多くを除外し，最重要単語の多くを分類に利用する手法の開発に取り組んだ．

正規またはスパムメールの片方のみに着目したとき，学習用と分類用

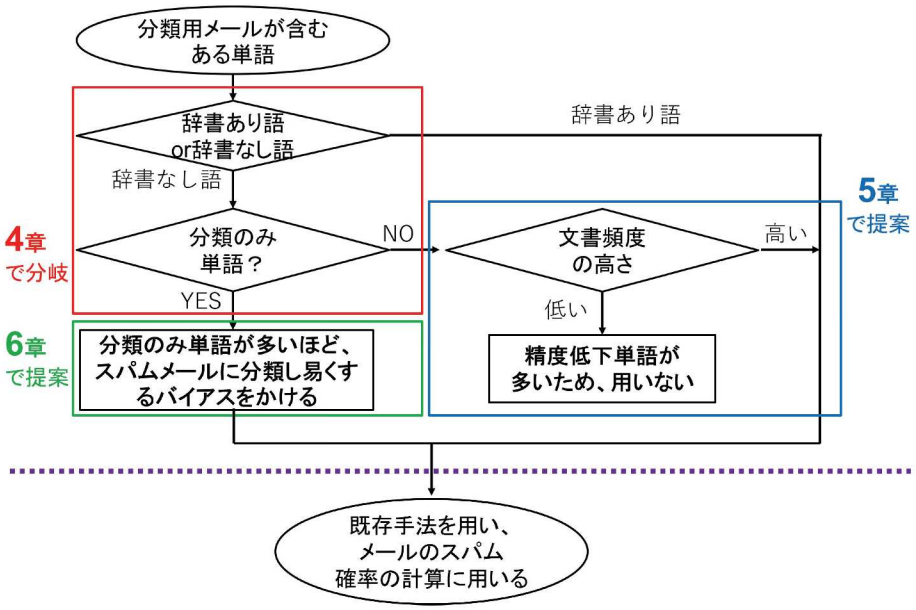

図 7.1: 辞書なし語の利用を既存フィルタリング手法に適用する流れ図

の両方に継続して出現するパターンが最重要単語であり，どちらかにのみ出現するパターンが精度低下単語である．これらの特徴の違いについて検証するため，単語ごとに文書頻度を調べた結果，最重要単語よりも，精度低下単語の文書頻度が低い傾向にあることを確かめた．

この結果より，文書頻度にしきい値を定め，これを超える語のみを分類に用いる手法を提案した（図 7.1 の青枠部）．これによって得られる効果について，bsfilter を用いて実験的に調べ，分類性能が向上することを確かめた．

6 章では，辞書なし語のうち，分類のみ語について，2.2.4 で述べた bsfilter での処理よりも，分類性能が向上する利用手法の開発に取り組んだ．まず，メールの件名と本文が含む単語の特徴の時間的変化に着目した分類実験を bsfilter を用いて行った結果，学習用メール群からの時間経過による分類性能の低下の原因は，スパムメールによるものが大きいことを確かめた．

この原因について調べるため，メールの件名と本文が含む単語に占め

る，分類のみ単語の割合を調べた．その結果，学習用メール群からの時間経過とともに，辞書なし語の分類のみ単語がスパムメールで増加傾向にあることを確かめた．

この傾向を分類に利用するため，分類対象のメールに辞書なし語の分類のみ単語が多く出現するほど，スパムメールに分類しやすくするバイアスをかける手法を提案した（図 7.1 の緑枠部）．

これによって得られる効果について，bsfilter を用いて実験的に調べた結果，辞書なし語の分類のみ単語にスパム確率 0.7 を設定することで分類性能が向上することを確かめた．

本論文が提案した処理は，図 7.1 の紫点線部よりも上，すなわち既存手法による処理よりも前に行うため，フィルタリング手法に依存せず，多くの手法に適用することができ，分類性能の向上が得られることが期待できる．

今後，この処理を実行するシステムを構築し，併用する既存手法や，扱うメールデータセットとの相性も含め，辞書なし語の利用の有効性について，さらなる検証を重ねていく．

現在高性能な分類が可能となっている既存手法の分類性能を，より完全なものに近づけるためには，フィルタリング手法の改良のみならず，その処理の前段階である単語の扱い方についても，新しい観点による改善を行う必要がある．本論文が提案する辞書なし語の利用は，その新しい手法の一つを与えるものである．

謝辞

山口大学創成科学研究科教授　松野浩嗣先生には，本当に感謝してもしきれない程, 感謝しております.

筆者が山口大学の学部生のとき，4回も留年しているのにもかかわらず，快く研究室に迎え入れていただきました．当時，筆者は複数のアルバイトを掛け持ちしておりまして，その内容や私の現状について，先生は親身になって話を聴いてくださいました.

そのとき，先生から「それが君の個性なんだ」という非常に前向きな言葉を言っていただけたことで，留年を繰り返して後ろ向きになっていた心境がものすごくラクになったことを，いまでも鮮明に覚えております．私が大学院進学についてご相談したときにも，ご快諾いただきまして，非常に感謝しております.

日々の研究打合せでは，ご多忙にもかかわらず，終始熱心なご指導をいただきました．ときには，厳しいと思ってしまったこともありましたが，私は，そこに愛情があったことを常に感じており，研究活動に励むことができました.

国際会議や研究会，非常勤講師などの貴重な機会も多くいただきました．先生から，聴き手に伝わりやすいプレゼン発表のやり方をご指導いただけたおかげで，研究会で優秀賞をいただけたり，非常勤講師で高い評価をいただくことができました.

先生は，どんなに疲れているときにも，手を抜くことなく，懇切丁寧にご対応くださり，論文作成のご指導や文章の添削などにもあたっていただきました．そのまっすぐな研究者としての姿勢と心構えは，筆者の目標であり，目指す研究者像となっております.

山口大学国際総合科学部教授　杉井学先生にも，研究の初期段階から，多くのご指導とご教示をいただき，メールフィルタリングの知識をご教授いただきました．筆者の研究の進捗が芳しくないときにも，優しく見

守ってくださり，丁寧なご指導をいただきました．心より深く感謝し，厚く御礼を申し上げます．

本論文をまとめるに際し，学位取得審査の副査として，本論文に対する真摯なるご検討と数多くの有益なご教示とご助言を受け賜りました，山口大学創成科学研究科教授　野崎浩二先生ならびに末竹規哲先生，ならびに同創成科学研究科准教授　浦上直人先生，ならびに同教育学部准教授　春日由美先生に厚く御礼を申し上げます．

野崎浩二先生には，私の進路についてのご相談にも親身に応じていただきました．筆者の兄とともに，多大なるご指導を受け賜りました．末竹規哲先生には，留年を繰り返して落ち込んでいた私に，優しく話しかけていただけたことで，授業に参加しやすくなり，大学にも行きやすくなりました．浦上直人先生には，修士課程での授業においてもお世話になりました．ソフトボール大会では同じチームとして，非常に楽しい時間を過ごすことができました．春日由美先生には，地域子育て支援拠点のスタッフが記述した業務日誌を解析する研究を通して，心理学の観点からも，数多くの有益なご教示とご助言を受け賜りました．深く感謝し，厚く御礼を申し上げます．

研究を遂行するにあたり，筆者が所属したネットワーク科学研究室の皆様には，多くのご協力をいただきました．毎日のように冗談を言い合い，非常に楽しい研究活動の日々を送ることができました．厚く御礼を申し上げます．

なお，本研究は，JST 次世代研究者挑戦的研究プログラム JPMJSP2111 の支援を受けたものです．ここに記して感謝の意を表します．

終わりに，筆者が繰り返した留年を許し，暖かく見守ってくださいました両親，ならびに，筆者を日々支えてくださいました兄と姉に，心から深く感謝しております．

最後に，筆者を支えて来られたすべての方々に，重ねて感謝を申し上げ，本研究の謝辞とさせていただきます．

令和 5 年 10 月
天満　誠也

参考文献

[1] 迷惑メール対策推進協議会, “迷惑メール白書 2021 年度版, ”一般財団法人 日本データ通信協会, Aug. 2021.

[2] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, “A Comprehensive Survey for Intelligent Spam Email Detection, ”IEEE Access, vol.7, pp.168261-168295, Nov. 2019. DOI:10.1109/ ACCESS.2019.2954791

[3] Paul Graham, “A Plan for Spam ”
http:// / www.paulgraham.com/ spam.html. ref. May. 23, 2022.

[4] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, G. Paliouras, and C.D. Spyropoulos, “An evaluation of naive bayesian anti-spam filtering, ” Proceedings of 11th European Conference on Machine Learning (ECML 2000), Barcelona, pp.9-17, Jan. 2000.

[5] E. Dada, J. S. Bassi, H. Chiroma, S. Abdulhamid, A. O. Adetunmbi, and A. E. Opeyemi, “Machine learning for email spam filtering: review, approaches and open research problems, ” Heliyon, vol.5, no.6, pp.1-23, Jun. 2019. DOI:10.1016/ j.heliyon.2019.e01802

[6] “Apache SpamAssassin, ” http:// / spamassassin.apache.org/ , ref. Dec. 2022.

[7] R. O. Duda, P. E. Hart and D. G. Stork, “Pattern Classification, 2nd Edition, ”New York, 2000.（尾上守夫（訳）“パターン識別 ”株式会社新技術コミュニケーションズ, 2001.）

[8] D. Sculley and G.M. Wachman, “Relaxed online SVMs for spam filtering,” Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp.415-422, New York, USA, Jul. 2007.
DOI:10.1145/ 1277741.1277813

[9] S. Shimozono, A. Shinohara,T. shinohara, S. Miyano, S. Kuhara and S. Arikawa, “KnowledgeAcquisition from Amino Acid Sequences by MachineLearning System BONSAI,” 情処学論, vol.35, no.10, pp.2009-2018, Oct. 1994.

[10] S. Usuzaka, KL. Sim, M. Tanaka, H. Matsuno and S. Miyano, “A Machine Learning Approach to Reducing the Work of Experts in Article Selection from Database: A Case Study for Regulatory Relations of S. cerevisiae Genes in MEDLINE,” Genome Inform, vol 9, pp.91-101, 1998.

[11] 杉井学, 松野浩嗣, “機械学習によるスパムメールの特徴の決定機表現,” 情処学研報, vol.2007, no.16, pp.183-188, Mar. 2007.

[12] Yao. X, Liu. Y, “A new evolutionary system for evolving artificial neural networks,” IEEE Transactions on Neural Networks, May. 1997.

[13] Y. Liao, V. R. Vemuri, “Use of K-Nearest Neighbor classifier for intrusion detection,” Computers & Security, vol.21, no.5, pp.439-448, Oct. 2002. DOI:10.1016/ S0167-4048(02)00514-X

[14] S.J. Delany, P. Cunningham, A. Tsymbal, and L. Coyle, “A case-based technique for tracking concept drift in spam filtering,” Knowledge-Based Systems, vol.18, pp.187{195, Aug. 2005.

[15] J. G. Cumming, “POPFile,” https://forest.watch.impress.co.jp/library/software/popfile/ , ref. Jan. 2023.

[16] Gary Robinson, “Spam Detection”
http://radio-weblogs.com/0101454/stories/2002/09/16/spamDetection.html. ref. May. 23, 2022.

[17] G. Robinson, “A statistical approach to the spam problem,” Linux Journal, vol.2003, no.107, Mar. 2003.

[18] “bsfilter,” https://ja.osdn.net/projects/bsfilter/, ref. Oct. 1, 2021.

[19] “Thunderbird,” https://www.thunderbird.net/ja/, ref. Oct. 1, 2021.

[20] 米倉正和，堀幸雄，後藤英一，“Support Vector Machineを用いた電子メールの自動分類,”情報処理学会研究報告，vol.31，pp.19-26, 2003.

[21] Deyue Deng，大塚孝信，伊藤孝行，“複数単語共起フィルタリングにより大規模化するデータを処理する有害文分類手法の提案，”情処学研報, vol.2013-ICS-171 no.2, pp. 1-8, Mar. 2013.

[22] 天満誠也，杉井学，松野浩嗣，“Jaccard係数を用いた単語の共起度に基づくメールフィルタの提案，”信学技報，vol. 118, no. 499, MSS2018-92, pp. 59-64, Mar. 2019.

[23] “MeCab: Yet Another Part-of-Speech and Morphological Analyzer,” https://taku910.github.io/mecab/, ref. Oct. 1, 2021.

[24] “NLTK,” https://www.nltk.org/, ref. Oct. 1, 2021.

[25] T. Fawcett “”In vivo” spam filtering: a challenge problem for KDD,” ACM SIGKDD Explorations Newsletter, vol.5, no.2, pp.140-148, Dec. 2003.
DOI:10.1145/980972.980990

[26] F. Fdez-Riverola, E.L. Iglesias, F. Díaz, J.R. Méndez and J.M. Corchado, “Applying lazy learning algorithms to tackle concept drift in spam filtering,” Expert Systems with Applications, vol.33, no.1, pp.36-48, Jul. 2007.
DOI:10.1016/j.eswa.2006.04.011

[27] 浦本直彦, “コーパスに基づくシソーラス-統計情報を用いた既存のシソーラスへの未知語の配置, ” 情処学論, vol.37, no.12, pp.2082-2189, Dec. 1996.

[28] P. Aguero, J. Moreira, M. Liberatori, J. Bonadero, J. Tulli, “Improving the Performance of Anti-Spam Filters Using Out-of-Vocabulary Statistics, ” Ingeniare, vol.17, no.3, pp.386-392, 2009.

[29] 小川健司, 稲葉宏幸, “記号と未知語の分布を用いたベイジアンスパムフィルタの提案”, 信学技報, vol. 108, no. 459, SITE2008-79, pp. 209-212, Feb. 2009.

[30] 平博順, 向内 隆文, 春野雅彦, “Support Vector Machine によるテキスト分類, ” 情報処理学会研究報告, vol.128, pp.173-180, 1998.

[31] 平博順, 春野雅彦, “Support Vector Machine によるテキスト分類における属性選択, ” 情処学論, vol.41, no.4, pp.1113-1123, Apr. 2000.

[32] “SpamAssassin public corpus, ” https://spamassassin.apache.org/old/publiccorpus/, ref. May. 23, 2022.

[33] T.A Meyer, and B Whateley, “SpamBayes: Effective open-source, Bayesian based, email classification system, ”CEAS 2004 - First Conference on Email and Anti-Spam, Mountain View, California, USA, Jul. 2004.

[34] “Spam Track, ” https://trec.nist.gov/data/spam.html, ref. Oct. 1, 2021.

[35] G.V. Cormack, “TREC 2007 spam track overview, ” Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, Nov. 2007.

[36] W. Pan ,J. Li, L. Gao, L. Yue, Y. Yang, L. Deng, and C. Deng, “Semantic Graph Neural Network: A Conversion from Spam Email

Classification to Graph Classification," Scientific Programming, vol.2022, no.11, pp.1-8, Jan. 2022.
DOI:10.1155/2022/6737080

[37] W. Liu, and T. Wang, "Online active multi-field learning for efficient email spam filtering," Knowledge and Information Systems, vol.33, no.1, pp.117-136, Oct. 2012.
DOI:10.1007/s10115-011-0461-x

[38] G.A. Miller, "WordNet: a lexical database for English," Communications of the ACM, vol.38, no.11, pp.39-41, Nov. 1995.
DOI:10.1145/219717.219748

[39] 永田昌明, 平博順, "情報論的学習理論とその応用：テキスト分類-学習理論の「見本市」-," 情報処理, vol.42, no.1, pp.32-37, Jan. 2001.

[40] Y. Yang and J.O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proceedings of the Fourteenth International Conference on Machine Learning(ICML-97), pp.412{420, Jul. 1997.

[41] H. Hofmann, H. Wickham and K. Kafadar, "Letter-Value Plots: Boxplots for Large Data," Journal of Computational and Graphical Statistics, vol.26, no.3, pp.469-477, Mar. 2017.
DOI:10.1080/10618600.2017.1305277