様式７号（第 12 条，第 31 条関係）
（様式７号）(Format No.7)　日本語版

学　位　論　文　要　旨
(Summary of the Doctoral Dissertation)

学位論文題目 (Dissertation Title)	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)
氏　名(Name)	天満　誠也(Seiya Temma)

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．その原因の一つに，スパム送信者がフィルタすり抜けのために，メール中の単語を記号，スペース及び HTML タグ等の組み込みによって改変する行為による影響がある．例えば，スパムメールには「price$ for be$t drug$!」，「priceＣＩＡＬＩＳ」，「sexual」等の文字列が含まれていることがあり，これらは，記号等の組み合わせの変更により，日々新しいもの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学習し，その結果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は，日々置き換えられているため，過去に受信したメール群が含まず，機械学習ができないものも多数ある．このような分類用メール群にのみ出現する文字列の特徴も分類に利用すれば，さらなる分類性能の向上が期待できる．

本研究では，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ものと，分類用メール群にのみ出現するものに区分したうえで，各々を分類に利用するための手法を開発することで，既存手法の分類性能を，完全フィルタリングに近づけることを試みる．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，上記の他にも，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，広く使われているメールフィルタである bsfilter 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，その利用手法を開発すれば，得られる分類性能の向上効果も大きいことが期待できる．

(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個数をそれぞれ集計した．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ところ，これらの品詞と比べて，辞書にない語には，学習用と分類用メール群の両方に出現し，かつ正規またはスパムメールの一方にのみ出現する，すなわち最も分類に貢献する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，すなわち利用のために特別な処理が必要となる語もかなり多いことがわかった．これらの結果から，辞書なし語のう

(和文 2,000 字程度 / 英文 800 語程度)
(about 800 words)

ち，分類に貢献する語の多くを利用する手法，及び分類用メール群にのみ出現する語を利用する手法の両方を開発し，これらを組み合わせることで，既存フィルタリング手法の性能を，完全なものに近づけることができる可能性を見いだした．

(3). 辞書にない語について，(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（処理が必要となる語）の利用手法の両方を開発した．

（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，出現傾向が学習用と分類用メール群で逆となる，すなわち分類性能が低下する出現パターンの語に分け，出現頻度をそれぞれ調べた．その結果，分類性能が向上する出現パターンの語は，分類性能が低下するものよりも，出現頻度が高い傾向にあることがわかった．分類性能が向上する出現パターンの語を多く利用するため，出現頻度がある程度高い語のみを分類に用いる手法を開発し，それを既存手法に導入する実験を行った結果，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特徴を調べるため，正規とスパムメールの各々で，単語の種類数を集計して比較した．その結果，正規よりもスパムメールに多く出現する傾向にあることがわかった．この特徴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する実験を行った．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%に向上することを確かめた．

上記(A)，(B)を併用することで，辞書にない語について，分類性能が向上する語の多くを利用しつつ，分類用メール群にのみ出現する語についても有効活用できる．これは，フィルタリング手法を実行する以前の，辞書にない語の扱い方を改善するものであるため，多くの既存手法と併用することができる．

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，本論文は辞書にない語の利用という，その観点のひとつを与えるものである．

様式７号（第 12 条，第 31 条関係）
（様式７号）(Format No.7)　英語版

学　位　論　文　要　旨
(Summary of the Doctoral Dissertation)

学位論文題目 (Dissertation Title)	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)
氏　名(Name)	天満　誠也(Seiya Temma)

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”, “sexual”, 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, 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 improve 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 using it in combination with existing methods and confirmed that it 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. 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.

(様式９号)

学位論文審査の結果及び最終試験の結果報告書

山口大学大学院創成科学研究科

氏　　名	天　満　誠　也
審査委員	主　査：　松　野　浩　嗣
	副　査：　野　崎　浩　二
	副　査：　末　竹　規　哲
	副　査：　浦　上　直　人
	副　査：　春　日　由　美
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

【論文審査の結果及び最終試験の結果】

電子メールは、コミュニケーション手段として確立した当初から、スパムメールの存在はインターネットのトラフィックを不必要に増加させるとともに、詐欺や情報漏洩などを引き起こす悪意のある行為が社会問題となってきた。

現在利用されているフィルタリングシステムでは、分類手法としてベイジアンフィルタ、決定木、サポートベクトルマシンなどの機械学習が使われており、95%を超える高い正解率を達成しているが、5%未満と少ないものの誤分類メールがあり、これらへ対応できる手法の開発が課題となっている。

スパムメール送信者は、フィルタすり抜けのために悪意をもって、偽の句読点、記号、スペース、及び HTML タブなどを単語に挿入したり、時々置き換えたりしている。本論文は、このような辞書に載っていない語のメール本文中の出現特性を解析し、その結果を応用してフィルタリング性能をさらに向上させる手法を開発したものである。

データとしてはメールフィルタ研究によく使われている TREC と SpamAssassin を用い、まず、辞書にない語、名詞、動詞、及び形容詞について分類実験を行い、辞書にない語が最もよい分類ができることを確かめている。さらに、辞書にない語について、「(A) 学習と分類メール群の両方に出現するものについては、正規またはスパムメールのどちらかに偏っているものが多いこと」、「(B)学習または分類メール群の片方に出現するものも含めると、分類メール群のみに出現するものが多いこと」を見出したうえで、学習／分類、正規／スパムの組合せからなるパターン別に単語の出現傾向を調べ、辞書にない語の種類数は「(A) 正しい分類に寄与するパターン（学習・分類ともに正規（スパム））の語は、そうでない語より多いこと」と「(B) 分類用メール群のみに出現する語では、正規よりもスパムに多い傾向にある」ことを確かめている。

（様式第９号）

以上で得られた知見から、辞書にない語を活用したスパムメールの分類手法として、次の２通りを開発している。

方法(A)：メール群に現れる単語の各々について、その単語を含むメールの本数を求め、文書頻度とする。文書頻度の高い単語には正しい分類に寄与するものが多い傾向があるため、文書頻度の閾値を適切に定めて、それよりも高い値をもつ単語のみを分類に利用する。

方法(B)：分類メール群のみに現れる単語において、辞書にない語は正規メールよりもスパムメールに現れる傾向があるため、分類メールのみに現れる単語には高めのスパム確率を意図的に適用する。

これら２つの方法の有用性を bsfilter を用いた計算機実験によって確かめ、以下の結果を得ている。

方法(A)：文書頻度の閾値を１から９まで変化させてスパム確率の変化を観察し、２以上の閾値において、分類精度の向上を確認している。特に、実施した実験では、閾値５が最適としている。

方法(B)：分類メール中の辞書にない語の各々について、0.0 から 1.0 まで 0.1 刻みでスパム確率を適用して分類性能を比較したところ、確率 0.7 のときが最適となることを確かめている。

本論文が開発した手法は、機械学習によるフィルタリング処理の前に適用するものであるため、現在実用となっているメールフィルタシステムに大きく改変を施すことなく組み込むことが可能である。

以上により、本研究は独創性、信頼性、有効性、実用性ともに優れ、博士（理学）の論文に十分値するものと判断した。

審査の過程において、類似研究との分類精度の比較をすること、分類メールのみに出現する辞書にない語に意図的に設定するスパム確率を定量的に説明すること等が求められたが、これに対応するために行った計算機実験の説明を含め、十分な回答がなされた。

公聴会において、分類メール群の新旧によってスパム確率が変化することの考察、正規メールの誤分類を防ぐための対応法、完全フィルタリングに向けた更なる改良への展望などが尋ねられたが、いずれの質問についても発表者から的確な回答がなされた。

以上、論文内容及び審査会、公聴会での諮問応答などを総合的に判断して、最終試験は合格とした。

なお、主要な関連論文の発表状況は以下の通りである。

1) 天満誠也、松野浩嗣、メールフィルタリング精度向上のための辞書にない単語の特性解析及び活用方法の提案、電子情報通信学会論文誌、vol. J105-D, no.11, pp.691-699, 2022 年 11 月発行。
2) 天満誠也、松野浩嗣、メールフィルタリング性能向上のための未学習辞書なし語の効果、情報処理学会論文誌：数理モデル化と応用、vol.16, no.2, pp.1-10, 2023 年 10 月発行予定。