様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 ２５ 日現在

機関番号：１５５０１
研究種目：挑戦的萌芽研究
研究期間：2016～2019
課題番号：１６Ｋ１２４３８

研究課題名（和文）正規メール特徴を重視し単語属性に着目した高精度・高速フィルタリング手法の開発

研究課題名（英文）Development of the method for mail filtering focused on fetures of ham mails and word attributes

研究代表者
杉井　学（Sugii, Manabu）

山口大学・国際総合科学部・准教授

研究者番号：００３５９９１０

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：　電子メール（以下メール）中の単語ではなく、その属性情報によってスパムメールを分類するメールフィルタ手法の開発が目的であった。これまでの単純な単語の出現頻度という属性に加え、品詞ごとの出現頻度や複数の単語が共出現する頻度（Jaccaard係数）、また一般的な辞書に存在しない単語のみの属性情報を用いることで、これまで以上にメール本文の特徴を捉えることができ、スパムメールを効率よく自動分類できることを明らかにした。

研究成果の学術的意義や社会的意義
　これまでのメールフィルタに用いられてきたメールを特徴付ける要素は、人が見た目で容易に判断できる単語そのものであったり、単純な単語の出現頻度であった。本研究課題での成果は、メールを特徴付ける新たな単語の属性情報を発見したことや、これまで注目されていなかった一般的な辞書に存在しない単語の属性情報の重要性を明らかにしたことなどである。特に後者は、メールフィルター研究において新たな展開をもたらすことにつながり、学術的意義は大きい。また、様々なメールフィルタシステムが開発される中で進む、スパムメール流通量の増加を鑑みれば、ネットワーク資源の効率的かつ適切な利用に向けた改善策としての社会的意義は大きい。

研究成果の概要（英文）：　Our goal was going to be development of the new method for mail filtering with not words but attributes of words in mail body. We made clear that the attributes (a word frequency of some parts of speech, a frequency of co-occurrence of some words; Jaccard index, and these attributes of only no dictionary words) can characterize and classify spam mails better.

研究分野：情報学

キーワード：メールフィルタ　スパムメール　属性情報　Jaccard係数　機械学習

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

(1)　様々なメールフィルタシステムが開発されているにもかかわらず、現在でもインターネット上を流通する迷惑電子メール（スパムメール）は、全体の 95%以上を占める。新たなスパムメールの対策を講じれば、スパムメール送信者はそれをかいくぐる新たな方法で送信を繰り返し、「いたちごっこ」の様相を呈している。

(2)　我々はこれまでに、機械学習システム BONSAI [1]を用い、文献の効率的収集システム [2]やゲノム情報からの重要配列パターン発見 [3]などを行ってきた。これらの研究と並行して、BONSAI に正規メール集合とスパムメール集合を与えて決定木抽出を試みたところ、スパムメールの特徴よりも正規メールの特徴を捉えて、99%以上の精度でメール分類ができる大変興味深い決定木が得られた。ここで、BONSAI にはメールをそのまま入力しているのではなく、メールに現れる単語を単語の出現頻度によって記号に置き換えてから入力している。つまり、単語の出現頻度などの属性情報によってメールをフィルタするための特徴づけをできることが示唆された。

２．研究の目的

(1)　メール中の単語ではなく、その属性情報によってスパムメールを分類するメールフィルタ手法を開発する。単語そのものではなく、属性情報のみを利用することで、記述言語を選ばないフィルタにすることや従来技術より少ない情報量で同等以上の分類精度を実現し、同時に処理時間の短縮などを試みる研究である。

(2)　正規メールの特徴を重視して分類パターンを抽出する機械学習システムによって、正規メールを選び出すことでスパムメールの複雑化・多様化の影響を受けないフィルタを実現し、同様に単語の属性情報を入力した従来システムの併用によって、これまでとは視点の異なる特徴抽出を実現し、それぞれの短所を補う高性能スパムメールフィルタの設計を目的とする。

３．研究の方法

(1)　正規メールの特徴を重視して分類パターンを抽出することができる BONSAI とベイズフィルタとして一般的な bsfilter [4]を用いる。それぞれのシステムに単語の属性情報を用いることで、言語非依存のシステム構築が可能であることを検証する。また同時に、言語論的な文章特徴抽出に関して過去の文献等を調査し、メールフィルタに最適な属性の発見を試みる。

(2)　従来手法である bsfilter に、属性情報を入力した時の分類精度と処理時間等を検証し、より高速なフィルタリングの可能性を探る。最後に、機械学習 BONSAI と bsfilter を組み合わせた併用フィルタを設計し、それぞれの短所を補完することができるパラメータ設定などを探る。得られた成果を元に、併用スパムメールフィルタのプロトタイプを作製し、性能評価試験をおこなう。

４．研究成果

(1)　メールフィルタのためのメール本文の特徴づけのために、従来の単語の出現頻度加え、下記の式(1)で得られる二つの単語の共出現頻度（Jaccard 係数）に注目し、Jaccard 係数を用いた単語間の共起ネットワーク(Figure 1)を KH Coder3 [5]および研究チーム構成員が受信したスパムメールを用いて作成して分析した。

$$Jac(A,B) = |A \cap B| / |A \cup B| \quad , \quad 式\ (1)$$

その結果、単語は共出現する頻度の高い単語同士がネットワークエッジで結ばれた Figure 1 のようないくつかのクラスタを作ることが分かった。さらに詳細分析を進めるために、Figure 1a 内のクラスタ A を構成する単語を取り除いて作成した共起ネットワークが Figure 1b である。ここには同様の内容のメールから抽出された単語で構成されるアルファベット B～E で示すクラスタが生じていることが分かり、単語の共起ネットワークを分析することで、メールの内容による分類がある程度可能であることが示唆された。

(2)　Figure 2 は Figure 1 と同様の分析を、正規メールについて行った結果である。やはり同様

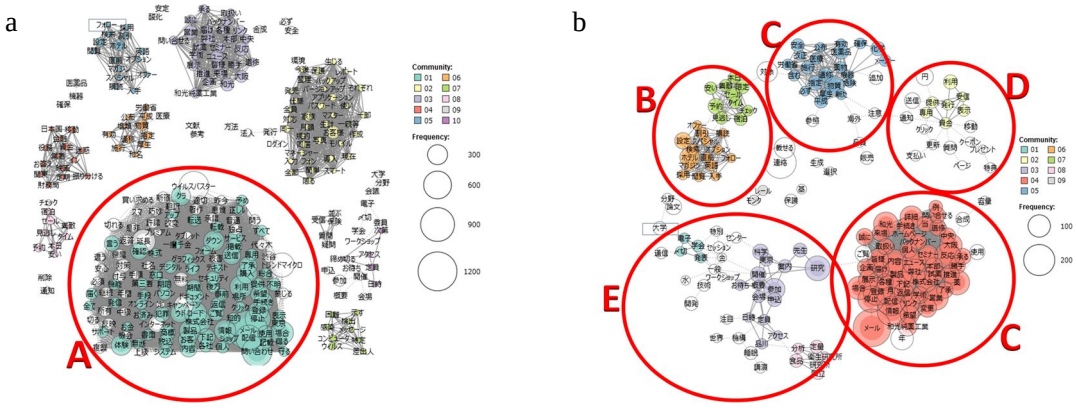

Figure 1　Jaccard 係数を用いたスパムメールに含まれる単語の共起ネットワーク図

の内容のメールに含まれる単語群がクラスタを作っているが、スパムメールと比べるとより受信者個人に関連する内容の単語（講義関連、研究関連、趣味など）が多い傾向にある。

(3)　単語間の共出現頻度（Jaccard 係数）を多次元尺度法を用いて、クラスタ分析を行った結果が、Figure 3 である。共起度合いを示す Jaccard 係数が低いために、Figure 1 では視覚化されていないが、多次元尺度法では新たなクラスタ（アルファベット F で表す集団）が抽出された(Figure 3a)。さらにクラスタ F に含まれる単語のみを取り出して、同様に多次元尺度法でネットワーク図を描いたものが Figure 3b で、このクラスタに含まれる単語は、いわゆる“出会い系メール”の勧誘を行うスパムメールから抽出されたものであることが分かった。

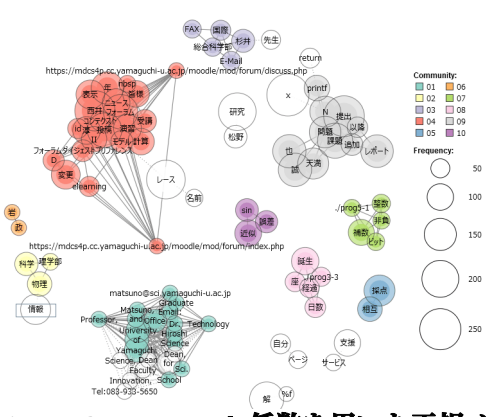

Figure 2　Jaccard 係数を用いた正規メールに含まれる単語の共起ネットワーク図

(4)　これらの結果は、二つの単語間の共起頻度（Jaccard 係数）を属性値として、メールフィルタに応用できる可能性があることを示唆している。また、出会い系メールのような、特徴的に共起する単語が含まれにくい（Jaccard 係数が低い）集団が存在することが明らかとなった。しかしながら、多次元尺度法などの分析手法を変えることでクラスタ化することができることも明らかとなった。

(5)　(1)〜(4)の結果を基に、下記の式(2)および(3)で求められる Jaccard 係数を用いた新たな単語の属性値（$JacDev(w_i, w_k)$）とメールの個性度（$DSI(d_i)$）を設定した。

$$JacDev(w_j, w_k) = \frac{Jac_H(w_j, w_k) - Jac_S(w_j, w_k)}{Jac_H(w_j, w_k) + Jac_S(w_j, w_k)}, \qquad 式\ (2)$$

$$DSI(d_i) = \frac{\sum_{j=1}^{T(d_i)-1} \sum_{k=j+1}^{T(d_i)} JacDev(w_j, w_k)}{{}_{T(d_i)}C_2}, \qquad 式\ (3)$$

Figure 4 は、2007 TREC Public Spam Corpus [6]をメールサンプルに用い、各学習に用いたメールの個性度およびスパム確率の分布を示している。また、縦軸が個性度またはスパム確率、横軸が各メールの番号を示している。メール番号については個性度またはスパム確率を基に昇順または降順にソートしている。

その結果、bsfilter で算出されるスパム確率を用いてメールを判定するよりも、個性度を用いてメールを判定したほうが、精度よくメールのフィルタができることが明らかになった（Table1）。

(6)　個性度とスパム確率の値に相関がないか確かめた。Figure 5 は縦軸をスパム確率、横軸を個性度として、学習に用いたすべてのメールの分布を示したものである。強い相関は見られなかったことから、二つの指標は異なる尺度でメールの特徴を表していることを示しており、これらの指標を用いたメールフィルタシステムの併用で、より精度の高いメールフィルタリングが実現できることを示している。

Table 1　学習メールの個性度とスパムメール確率による分類精度比較

	Proposed method		bsfilter	
	Ham	Spam	Ham	spam
Precision (%)	99.98	99.98	99.94	99.94
Recall (%)	99.98	99.98	99.94	99.94
F value	0.9998	0.9998	0.9994	0.9994

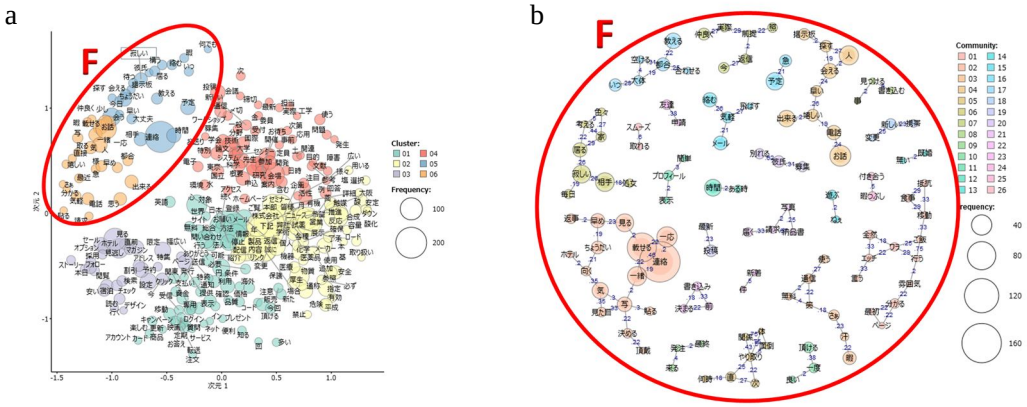

Figure 3　Jaccard 係数を多次元尺度法でクラスタリングした共起ネットワーク図

a：個性度の分布　　b：スパム確率の分布

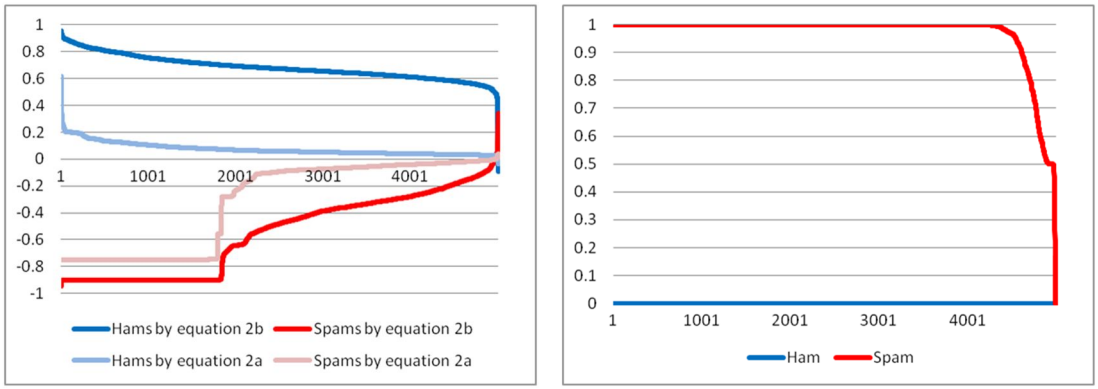

Figure 4　各メールの個性度（DSI）およびスパム確率の分布

a：正規メールの分布　　b：スパムメールの分布

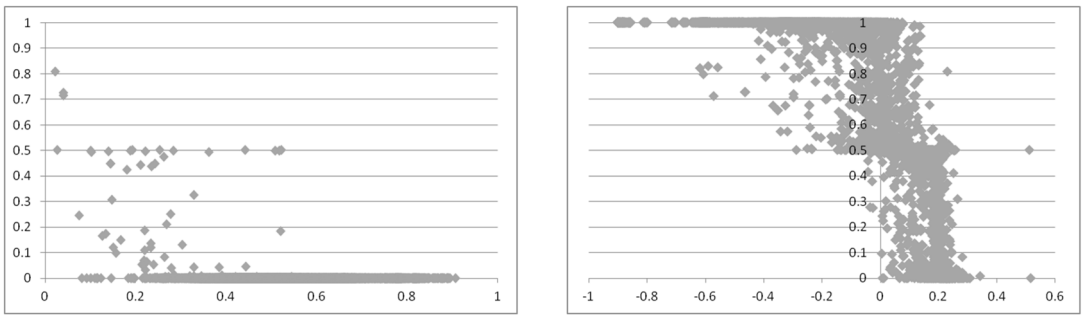

Figure 5　個性度とスパム確率の相関

＜引用文献＞

[1] S. Shimozono, A. Shinohara, T. Shinohara, S. Miyano, S. Kuhara , S. Arikawa, “Knowledge acquisition from amino acid sequences by machine learning system BONSAI,” Trans. Information Processing Society of Japan, 2994.

[2] S. Usuzaka, L. Kim, M. Tanaka, H. Matsuno , 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 Ser Workshop Genome Inform, 1998.

[3] C. Miyakawa, M. Sugii, M. Hiroshi , S. Miyano, “Computational predictions for functional proteins working after cleaved in apoptotic pathway,” Proc. International Conference on Complex, Intelligent and Software Intensive Systems (CISIS2009), 2009.

[4] bsfilter, 05 06 2020. [オンライン]. Available: https://ja.osdn.net/projects/bsfilter/.

[5] 樋口　耕一, “KH Coder：計量テキスト分析・テキストマイニングのための不フリーソフトウェアー,” 5 6 2020. [オンライン]. Available: https://khcoder.net/.

[6] 2007 TREC Public Spam Corpus, 5 6 2020. [オンライン]. Available: https://plg.uwaterloo.ca/~gvcormac/treccorpus07/.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　4件）

１．発表者名
Seiya Temma, Manabu Sugii, Hiroshi Matsuno

２．発表標題
Searching Attribute Information for Mail Filtering based on Text Mining

３．学会等名
The 33rd International Technical Conference on Circuit/Systems Computers and Communications（国際学会）

４．発表年
2018年

１．発表者名
天満誠也、杉井　学、松野浩嗣

２．発表標題
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

３．学会等名
電子情報通信学会　システム数理と応用研究会（MSS）

４．発表年
2019年

１．発表者名
Manabu Sugii, Nozomi Fujii, Hiroshi Matsuno

２．発表標題
An Effect of Word Order for Mail Classification by Bayesian Method

３．学会等名
The 32th International Technical Conference on Circuit/Systems Computers and Communications（国際学会）

４．発表年
2017年

１．発表者名
天満　誠也、杉井　学、松野　弘嗣

２．発表標題
メールフィルタのためのテキストマイニングを用いた属性情報の探索

３．学会等名
電子情報通信学会　システム数理と応用研究会

４．発表年
2018年

１．発表者名
藤井　望、杉井　学、松野　浩嗣

２．発表標題
ベイジアン方式メールフィルタリングにおける変換単語の属性情報探索

３．学会等名
電子情報通信学会　システム数理と応用研究会

４．発表年
2017年

１．発表者名
Seiya Temma, Manabu Sugii, Hiroshi Matsuno

２．発表標題
The Document Similarity Index based on the Jaccard Distance for Mail Filtering

３．学会等名
The 34nd International Technical Conference on Circuit/Systems Computers and Communications（国際学会）

４．発表年
2019年

１．発表者名
Manabu Sugii, Seiya Temma, Hiroshi Matsuno

２．発表標題
Extracting Co-occurrence Feature of Words for Mail filtering

３．学会等名
The International Conferrence on Artificial Life and Robotics（国際学会）

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	松野　浩嗣 (Matsuno Hiroshi) (10181744)	山口大学・大学院創成科学研究科　・教授 (15501)	