2E1-04

# Self Organizing Map に基づく 電子メール分類ルール自動生成手法及び評価

## An Method and Experimentation of Rule induction for E-Mail Classification based on a Self Organizing Map

平岡佑介[*1] 大囿忠親[*1] 伊藤孝行[*1] 新谷虎松[*1]
Yusuke Hiraoka Tadachika Ozono Takayuki Ito Toramatsu Shintani

[*1]名古屋工業大学　大学院工学研究科情報工学専攻
Graduate School of Univercity, Nagoya Institute of Technology

In this paper, we describe automatic rule induction for e-mail classification. E-mails have been used generally, and users use e-mails for knowledge collection indeed communication and management. Users receive many kind of e-mails, and it is very bother to sort by hand. To support users' e-mail sorting many systems have been proposed. These systems need many pre-classified mails to learn to classify. Users have to compose pre-classified e-mails. Also, There is a problem that classified results are difficult for users to use. We have implemented a rule induction system for e-mail by using clustering results of e-mails based on content of e-mails. For details, we have focused that deviation of self organized mail map. Finally, we propose experimental method to evaluate our system, and show the experimental results.

## 1. はじめに

本稿では，メールボックス自動生成システムの提案および本システムを実現するための電子メールマップ生成手法及び分類ルール生成手法を示す．電子メールが一般的に利用されるようになり，連絡手段だけでなく，知識及び情報の収集に電子メールが利用されるようになった [1]．知識及び情報の収集を目的として電子メールを利用する際，ユーザは興味のある多くのメーリングリストに登録し，大量の電子メールを受信する．通常，ユーザはフォルダを用いた電子メール分類を行うことが可能である．しかし，ユーザが人手でフォルダへ電子メールを分類することは多くのコストが必要となる．そのためメーラでは，大量の電子メールを整理するため，ユーザに電子メールの分類ルールの作成を行わせることで本問題を解決している．しかし，類似する内容のメーリングリストの電子メールを１つのメールボックスへ分類したい場合等は複数の分類ルールを組み合わせる必要があり，ユーザにとって負担になる．本研究では，Self Organizing Map[2] (以下 SOM) を用い複数の電子メールを２次元上に配置したマップを生成する．本マップの特長として互いに類似する電子メールはマップ上で近くに配置され，互いに異なる電子メールはマップ上でより離れた位置に配置される．配置されたメールの分布に着目し，電子メール分類ルールの生成を行う．本機能を備えた電子メール閲覧支援システムとして，メールボックス自動生成機能を持つ電子メール閲覧支援システム WisdomMail を試作した．

以下に本稿の構成を示す．第２章では，本研究の関連研究を示す．第３章では，本研究で開発を行った電子メール閲覧支援システム Wisdom Mail について述べる．次に，第４章で Self Organizing Map を用いた電子メール分類手法の流れについて述べ，実際に電子メールを分類した結果を示す．第６章で電子メール分類結果を用いて実際に分類ルールを生成する手法について述べ，第７章で評価，考察を行う．最後に第８章で本稿をまとめ，本研究の今後の課題について述べる．

連絡先: 平岡佑介，名古屋工業大学大学院工学研究科情報工学専攻，466-8555 愛知県名古屋市昭和区御器所町，hiraoka@ics.nitech.ac.jp

## 2. 関連研究

以下に本研究の関連研究を示す．あらかじめユーザに電子メールを分類させることで分類器を学習し，電子メールの自動分類を実現する研究として POPFile[*1]及び Crawford の研究 [3] がある．Crawford[3] はあらかじめユーザに分類させた電子メールを用い，自動的に電子メールを分類するルールの生成を行うシステムを提案している．POPFile はユーザが新着メールを分類し，分類結果を用いてベイジアンネットワークを用いた分類器を学習し，自動的な分類器の生成を行っている．これらのシステムでは，誤分類の際，ユーザが正しいメールボックスへ電子メールを分類し直す必要があり，ユーザに負担が必要となる．また，SwiftFile[4] では，新規メールを受信した際に TF-IDF に基づくテキスト分類器を用い，新規メールの分類先メールボックスとして３つの候補を出力する．ユーザは新着メールが候補となった３つのメールボックスから最も最適なメールボックスを選ぶ．分類先を３つにしぼるため，ユーザは容易に新着メールを適切なメールボックスへ移動することができる．以上の流れによってインクリメンタルな分類器の学習を行っている．しかし，これらのシステムでは，分類先のメールボックスはあらかじめユーザが作成しておく必要がある．また，ユーザに電子メールを分類させる必要があり，ユーザにとって負担が必要となる．

また，電子メールをクラスタリングすることにより，電子メールの自動分類を行う研究として，ACEMS[5] がある．ACEMS では新着メールのクラスタリングを行った上で，各クラスタへラベル付けを行いユーザに提示することでユーザの新着メール処理の支援を行っている．しかし，分類された結果およびラベル付けはユーザにとってわかりにくいという問題点がある．

本研究は，既存の電子メール分類と異なり，電子メールをクラスタリングした結果から電子メール分類ルールを生成する点で以上の研究と異なる．

*1 POPFile:http://popfile.sourceforge.net/

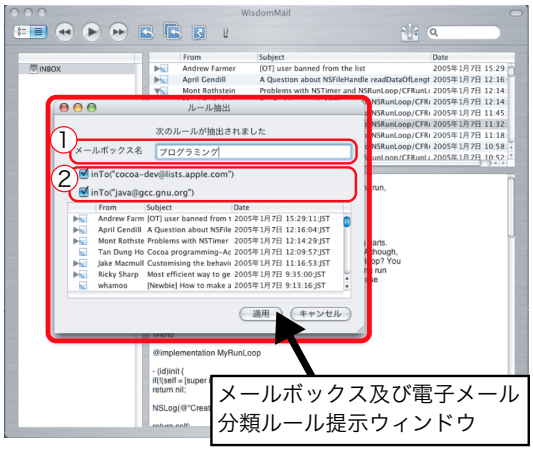


図 1: WisdomMail のスナップショット

## 3. 電子メール閲覧支援システムのインタフェース

図 1 に本研究で開発を行った電子メール閲覧支援システム WisdomMail のスナップショットを示す．本システムでは，未分類の受信メールはすべてメールボックス INBOX に保存される．本システムは一定時間毎に INBOX に含まれる電子メールを分類ルール自動生成機構に入力し，出力された新規メールボックス及び電子メール分類ルールをポップアップウィンドウを用いてユーザに提示する．電子メール分類ルールは複数の条件の組み合わせで作成されており，ユーザが変更及び管理を行うことができる．図 1 の例では，図中の (2) に，電子メールの宛先に”cocoa-dev@lists.apple.com”が含まれる．または，”java@gcc.gnu.org”が含まれる電子メールを同一のメールボックスに分類するためのルールが示されている．ユーザが生成されたルールを見て図中の (1) に示すテキストフィールドへメールボックス名を入力し，適用ボタンを選択することによって生成された分類ルールが適用される．一度，電子メール分類ルールが生成されるとそのルールを満たす新着メールはすべて生成されたメールボックスへ移動する．

## 4. SOM に基づく電子メールマップの生成

### 4.1 Self Organizing Map（SOM）

SOM は教師なし競合強化学習及び近傍学習により，ある分布に従う n 次元ベクトルで表現された入力データの特徴を抽出し，その分布を近似した特徴マップを生成する．SOM のネットワークは，入力層と 2 次元平面マップ上にノードを格子上に配置した出力層の 2 層からなり，入力されたデータが出力ノードの 1 つに出力される．各ノードは 2 次元平面上に表記されるため，類似した特徴を持つデータはマップ上の近い位置に出力される．生成されたマップはそれぞれのデータの位置関係によって類似しているデータかどうかが直観的に理解しやすいという点からデータの視覚化に利用できる．また，データの出力された位置を用いることでデータの分類を行うことが可能である．本研究では，SOM を用いた処理を行うために SOM_PAK[*2]を用いた．

*2 http://www.cis.hut.fi/research /sompak/

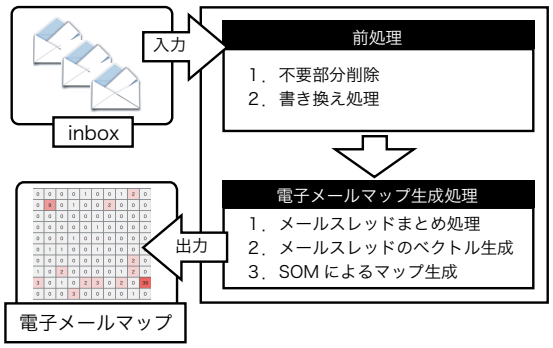


図 2: Self Organizing Map に基づくメール分類手法の流れ

### 4.2 電子メールマップの生成

図 2 に Self Organizing Map に基づく電子メールマップ生成手法の流れを示す．まず，前処理として電子メール本文中の記号及び署名及び広告といったマップ生成処理を行う上で必要でない部分の削除を行う．次に互いに返信関係のある電子メールをスレッドとしてまとめ，スレッドに対して分類処理を行う．現在の実装では，「Re:」を除いたサブジェクトが同一のメールを返信関係を持つメールとして 1 つのスレッドにまとめた．また，スレッドのサブジェクトとしてスレッドが開始された電子メールのサブジェクトを用い，スレッドの本文にはスレッドに含まれる全ての電子メール本文を結合した文字列を用いた．次に，得られたスレッドの特徴ベクトル表現を取得する．ベクトル表現への変換処理を以下に示す．

- **ステップ 1** スレッド本文又はサブジェクトに単語 w が出現するスレッドの数を全て調べる．電子メールの本文から単語を抽出するために TreeTagger[*3]を用いた．TreeTagger を用いて形態素解析した結果から SMART システム [6] で利用されている不要語を除いた単語を利用した．
- **ステップ 2** 単語を出現するスレッド数の多い順に並べ替えて，上位 N 個の単語 $(w_1, ..., w_N)$ を特徴ベクトルの要素とする．
- **ステップ 3** スレッドの特徴ベクトルの n 番目の要素 $f_n$ を次式で求める．

$$f_n = \begin{cases} 1 & \text{サブジェクトに単語 } w_n \text{ が含まれる} \\ 0.5 & \text{本文にのみ単語 } w_n \text{ が含まれる} \\ 0 & \text{それ以外} \end{cases}$$

得られたベクトル表現を SOM 分類器に入力し，分類処理を行う．本システムの現在の実装では 10 × 10 の SOM を用いている．

### 4.3 SOM による電子メール分類の特徴

実際に分類処理を行い，2 次元平面に視覚化した結果を図 3 に示す．本例では実験として，人工知能学会のメーリングリスト中から英文で書かれた電子メール，Java に関するメーリングリストである Cocoa Dev List[*4]，及び GCJ Java Milinglist[*5]を含めた 150 通のメールを対象に分類処理を行った．図中の格

*3 http://www.ims.uni-stuttgart.de/projekte/corplex /TreeTagger/DecisionTreeTagger.html
*4 http://lists.apple.com/mailman/options/cocoa-dev/
*5 http://gcc.gnu.org/ml/java/

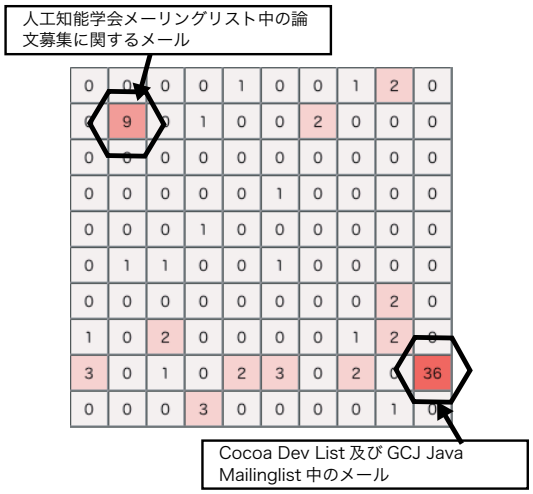


図 3: SOM による電子メール分類例

表 1: 電子メール分類ルールにおける条件

| 条件 | 説明 |
|---|---|
| inSubject(w) | 電子メールのサブジェクトに語 w が含まれる |
| inFrom(add) | 電子メールの From に宛先 add が含まれる |
| inTo(add) | 電子メールの To に宛先 add が含まれる |
| inCc(add) | 電子メールの Cc に宛先 add が含まれる |
| inCommunity(add) | 電子メールの From, To, または Cc のいずれかに宛先 add が含まれる |

子上に分かれた各ノードがマップの各ノードを表しており，各ノード中の数字はそのノードへ出力されたメールスレッドの数を表している．分かりやすさのため，出力されたメールスレッドの数が多いノードほど濃い色を用いて表示した．図中のマップにおいて丸で囲った部分に含まれるノードにそれぞれ，人工知能学会のメーリングリストの電子メール及び Java に関するメーリングリストの電子メールが出力され，本結果から 2 つのトピックに関する電子メールを分類できたと考えられる．本分類結果から類似する電子メールが同一のノードに出力される傾向があることがわかった．

## 5. 電子メール分類ルール生成手法

### 5.1 電子メール分類ルールの定義

電子メール分類ルールとは，電子メールを指定のメールボックスへ移動するためのルールであり，1 つ以上の条件の AND または OR で表現される．表 1 に本システムで利用した条件の一覧を示す．本ルールに従い，人工知能学会のメーリングリスト[*6]から Call for Paper に関するメールをメールボックス Mailbox1 へ移動するルールを記述した例を次に示す．

inFrom(”admin@ai-gakkai.or.jp”) ∧
inSubject(”CFP”) → move(”Mailbox1”)

*6 http://www.ai-gakkai.or.jp/jsai/ml/

表 2: 実験対象メール

| メーリングリスト | 内容 | メール数 |
|---|---|---|
| 人工知能学会のメーリングリスト | 論文募集，教官募集 | 50 通 |
| Cocoa Dev List | プログラミング | 50 通 |
| Gnu Java Milinglist | Java プログラミング | 50 通 |
| The Cygwin Project mailing list[*7] | UNIX 環境 cygwin | 50 通 |

### 5.2 電子メール分類ルール生成処理

SOM によって電子メールを分類した結果を用いてルールの生成を行う．ルールの生成は同一のノードに出力されたメールスレッドの満たす条件を手がかりに行う．以下に OR 結合したルールの生成処理を示す．

1. マップ中のある (x,y) 座標に分類されたスレッドのリスト $T_{xy}$ を取得する．
2. $t_{xyn}(t_{xyn} \in T_{xy})$ の満たす条件 c を取得する．
3. 全メールスレッド中から c を満たすスレッドの数を N(c) とし，c を満たす $T_{xy}$ 中のスレッドの数を $n_{xy}(c)$ とする．
4. $n(c)/N(c) > \alpha$ のとき条件 c を OR 条件の一つとして採用する．現在の実装では，$\alpha = 0.3$ を用いた．

また，ルール生成の際に同一の条件を満たすルールが同程度出力されているノードは結合を行い，結合したノードに対してもルール抽出処理を行う．具体的には，ある条件 c を満たすスレッドの数 $n_{xy}(c)$ で並び替えた時に互いに隣り合うノードを結合する．

## 6. 評価実験

### 6.1 電子メールマップ生成実験

本システムでは SOM を用いて電子メールのマップを生成し，生成されたマップを利用して電子メール分類ルールを生成している．本研究では，予備実験として SOM を用いて複数のメーリングリストに送られて来た電子メールのマップを生成し，評価を行った．実験対象として，表 2 に示す電子メールを用いてマップ生成を行った．

実験の結果として出力されたマップを図 4 に示す．分かりやすさのため，各メーリングリスト中のメールマップを示す．各マップ中の各ノードの数字はそのノードに対象とするメーリングリスト中のメールスレッドが出力された数を示す．これらのマップの特徴として人工知能学会のメーリングリストを対象としたマップ中の (1) で示されたノードに人工知能学会のメーリングリスト中の CFP に関する電子メール 30 通のうち 36%の 11 通が出力された．また，cocoa dev list，Gnu Java Mailinglist 及び The Cygwin Project mailing list のメールを対象とした各マップ中の (2) 及び (3) で示された各ノードにそれぞれのメーリングリストのメールの 30%以上が出力されている．本結果により，人工知能学会のメーリングリスト及びソフトウェア関連のメーリングリストを分けるマップが生成されたことが分かる．

### 6.2 電子メール分類ルール生成実験

本システムによって出力される電子メール分類ルールの有用性を示すために分類ルール生成実験を行った．本実験では表 2 に示す電子メールを用い，図 4 に示す電子メールマップを用

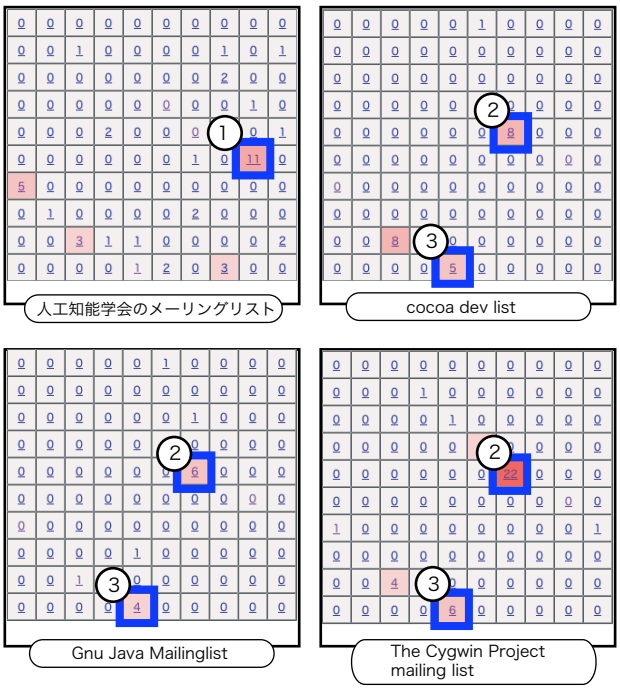


図 4: 生成されたマップ

表 3: 本システムによって生成された分類ルール例

| | ルール記述 |
|---|---|
| 1 | inFrom(”admin@ai-gakkai.or.jp”) ∧ inSubject(”CFP”) → move(”Mailbox1”) |
| 2 | inCommunity(”cocoa-dev@lists.apple.com”) ∨ inCommunity(”java@gcc.gnu.org”) ∨ inCommunity(”cygwin@cygwin.com”) → move(”Mailbox2”) |

いて分類ルールの生成を行った．表 3 に本システムによって生成されたルールを示す．表中のルール 1 は人工知能学会のメーリングリストから CFP に関する電子メールを分類するルールで，図 4 に示すマップ中の (1) で囲ったノードから生成された．また，ルール 2 は人工知能学会以外の 3 つのメーリングリストの電子メールを分類するルールで，図 4 に示すマップ中の (2) 及び (3) の 2 つのノードを結合した情報から生成された．

また，多数のメーリングリストを用いて分類ルールの生成を行う際に定量的に評価を行う指標が必要となる．ユーザが生成する分類ルールとシステムの生成する分類ルールの一致度を精度として用いることでシステムの評価を行う．被験者に実験に用いた各メーリングリストの電子メールを閲覧してもらい分類ルールを作成してもらった．被験者の作成した分類ルールと本システムの生成した分類ルールの精度を利用する．式 (1) に精度を示す式を表す．各式において $R_u$ はユーザの生成した分類ルールの集合を表し，$R_s$ は本システムの生成したルールの集合を表す．

$$\text{精度} = \frac{n(R_u \cap R_s)}{n(R_u)} \quad (1)$$

精度が高いほどユーザの作成したルールと同一のルールが出力され，本ルール生成機構がユーザにとって有用であることを示す．9 種類のメーリングリストの電子メールをそれぞれ 20 通ずつ用意し本手法で実験を行ったところ精度は 0.57 であった．ユーザの作成した分類ルールと違った分類ルールがシステムから生成された例を以下に示す．ユーザは Java プログラム言語に関する電子メールを分類するルールを生成したが，システムがそのルールを包含する任意のプログラム言語に関する電子メールを分類するルールを生成した．本問題点は，単一のノードに出力されている電子メールの満たす条件を分類ルールに採用するかの判定を行う閾値 $\alpha$ を調節することで解決できると考えられる．

## 7. まとめ及び今後の展望

本稿では，ユーザの電子メール分類ルール作成に必要となるコストの軽減を目的として，電子メール分類ルール自動生成手法について述べた．本システムは Self Organizing Map を用いて電子メールマップを生成し，生成されたマップの偏りを用いてメールボックス及び電子メール分類ルールの生成を行う．生成されたルールをユーザに提示することで，ユーザが電子メール分類を行う際の負荷の軽減を行うことができる．

以下に今後の展望を挙げる．現在の実装では，既に電子メール分類ルールが抽出されているメーリングリストの送信者のメールアドレス変更があった場合に対応することができない．送信者のメールアドレス変更等の手がかりの変更に対応できる電子メール分類ルールの生成を行う必要がある．本課題に対して，既に生成されたルールで分類された電子メールと未整理の電子メールを用いて本ルール生成処理を定期的に行うことによってメールアドレスの変更前及び変更後の電子メールを単一のメールボックスへ分類するためのルールを生成することができると考えられる．

## 参考文献

[1] Steve Whittaker, Candace Sidner: ”Email overload: exploring personal information management of email”, Human factors in conputing systems(1996)

[2] T.Kohonen, 徳高平蔵, 岸田悟, 藤村喜久郎:”自己組織化マップ”, シュプリンガー・フェアラーク東京 (1996)

[3] Elisabeth Crawford, Judy Kay, Eric McCreath : ”Automatic Induction of Rule for e-mail Classification”, In Proc. of the 6th Australasian Document Computing Symposium(2001)

[4] Richard B. Segal, Jeffrey O. Kephart : ”Incremental Learning in Swift File”, In Proc. of the 7th International Conference n Machine Learning(2000)

[5] Olle Balter, Candace L. Sidner: ”Bifrost Inbox Organizer: Giving users control over the inbox”, In Proc of the Second Nordic Conference on Human-Computer Interaction(2002)

[6] Salton G. : ”The SMART Retrieval System - Experiments in Automatic Document Processing”, Prentice Hall(1971)