5P-7

電子メールの定型性を用いた自己組織化に基づくスマートメールボックス自動生成

平岡 佑介 † 大囿 忠親 † 伊藤 孝行 † 新谷 虎松 †
名古屋工業大学大学院工学研究科情報工学専攻 ‡

1 はじめに

近年電子メールの利用が増加しており，電子メールを対象としたマイニングが盛んに行われている．本稿では，電子メールの自己組織化結果からのスマートメールボックス生成手法を提案する．スマートメールボックスとは電子メール分類ルールが付与されたメールボックスで，メールボックスに対応する分類ルールを満たす電子メールのリストが表示される．電子メール分類ルールはユーザにとって分かりやすく，分類ルールの修正を行うことで分類されている電子メールを調整することができる．本手法は，従来の単語の出現頻度に基づく電子メールの特徴化に加え，電子メール中のテンプレートの出現及び対象とする電子メールとテンプレートの類似度を用いた定型性を考慮した自己組織化における偏りに着目する．

2 電子メール特徴ベクトル

2.1 電子メール特徴ベクトルの概要

本手法では，電子メールの特徴ベクトルとして，互いに返信関係及び転送関係のある電子メールを統合した電子メールスレッドの特徴ベクトル $V_{thread}(M) = (k_1, ...k_n)$ 及び電子メールの定型性に基づく特徴ベクトル $V_{template}(M) = (j_1, ...j_s)$ を生成し，２つのベクトルを組み合わせたハイブリッドベクトル $V(M) = (k_1, ..., k_n, wj_1, ..., wj_s)$ を用いる．ただし，w は正規化を行うための重みである．$V_{thread}(M)$ の生成はそれぞれ対象とするスレッド中に出現する単語に基づいて生成する．具体的には，要素 k_n はメールスレッド中の単語 w_n の tfidf 値を用いる．$V_{template}$ の生成は第 2.2 節にて述べる．

2.2 電子メールにおける定型性

本研究では，直感的な定義として，定型性を”電子メール送信者が電子メールを書く際の目的に従って用いるテンプレート”とする．表１にテンプレートの例を示す．本研究ではテンプレートとして大別して２種類のテンプレートを用いる．特徴表現テンプレートは電子メール中のある段落を対象としたテンプレートとする．各特徴表現テンプレートは区切り文字^^で区切ることで類似したテンプレートを単一のテンプレートとして複数指定することができる．特徴表現１は日時情報と会場情報が含まれる段落が存在する電子メールを表す．ここで，<SEPARATOR>は区切り文字で，コロンなどの記号とマッチングする．<DATE>及び<PLACE>はそれぞれ，日付，場所とマッチングする．特徴表現２は電子メール中にプログラムソースが含まれるかどうかを示す．文例テンプレートは電子メール全文を対象としたテンプレートとする．例えば，文例１は備品購入報告のテンプレートを表す．テンプレートを用いた定型性に基づく電子メールスレッド M の特徴ベクトルの第 n 要素を以下に示す．ここで，各テンプレート T_n に複数のテンプレートが指定されている場合，i 番目のテンプレートを T_{ni} で表す．

表 1: 電子メールテンプレートの例

	テンプレート例
特徴表現１	日時<SEPARATOR><DATE><SEPARATOR> 会場<SEPARATOR><PLACE> ^^ 開催日時 <SEPARATOR> <DATE> <SEPARTOR> <TIME> <SEPARATOR> 開催場所 <SEPARATOR> <PLACE> ^^ ...
特徴表現２	<PROGRAMSOURCE>
文例１	<ANY>の購入を行いました． 購入日：<DATE> 品名：<ANY> 数量：<NUMBER>

テンプレート T_n が特徴表現テンプレートの場合 各テンプレート T_{ni} に対して，M が T_{ni} のうちいずれかを満たす場合 $j_n = 1$ とし，M が T_{ni} を全て満たさない場合は $j_n = 0$ とする．

テンプレート T_n が文例テンプレートの場合 電子メールとテンプレート T_n との類似度を用る．具体的には $j_n = sim(M, T_n)$ を用いて要素 j_n を決定する．ここで，$sim(M, T_n)$ は電子メール M 及びテンプレート T_n の出現単語の頻度を並べたベクトルの余弦を用いる．

3 スマートメールボックス生成手法

各スマートメールボックスは電子メール分類ルールを持ち，電子メールの満たすプリミティブ条件の AND (&で表す）または OR（|で表す）で表現される．プリミティブ条件の例として，Subject に語 w を含む inSubject(w) がある．同様に inFrom(add)，inTo(add)，inCc(add)，及び From，To，または Cc のうちのいづれかに宛先 add が含む inCommunity(add) を用いる．

本手法では，生成された電子メール特徴ベクトルを Self Organizing Map[1]（以下 SOM）に入力してマップの生成を行い，生成されたマップを用いてスマートメールボックス生成を行う．本電子メールマップ生成では，SOM_PAK[1]の距離関数に余弦を用いるように修正したライブラリを用いる．また，本手法では 10x10 の SOM マップを用い，ハイブリッドベクトルの要素として合計 600 次元のベクトルを用いた．

生成されたマップの例を図１に示す．本システムを用いて電子メールテンプレートを用い，電子メール 270 通に対して SOM マップ生成を行った．対象として，スクリプト言語 AppleScript に関する applescript-user，Java に関する cocoa-dev，gnujava，java-dev-apple，Core Java Technology，UNIX ターミナルを実現するアプリケーションに関する cygwin，アジア系言語と英語に関する翻訳に関する honyaku，情報技術に関する ITTips，

Automatic Smart Mailbox Generation based on Self-Organization by using a Template of an E-Mail

[1] http://www.cis.hut.fi/research/som_pak/

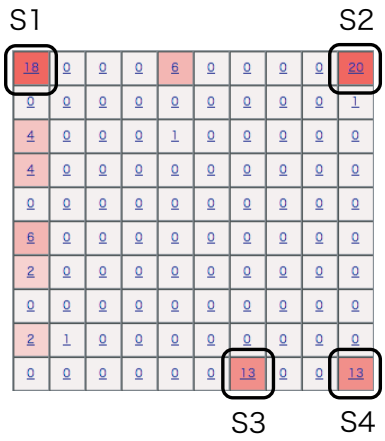

図 1: 電子メールマップ（定型性を考慮する）

及び会議情報や教員募集など研究者に関する jsai のメーリングリストを用いた．図に示した電子メールマップの各セルは SOM 上の 2 次元平面上に配置された各セルを示す．各セル中に書かれた数字はセルに出力されたスレッド数を示し，出力されたスレッドの多いセルほど濃い色で表示される．本図から，マップ中にいくつかのスレッドが多く出力されたセルがあることがわかる．また，図に示したマップは電子メールの定型性を考慮しないマップに比べ，スレッドの多く出力されたセルが 4 つ多く，定型性を考慮したマップのほうがより細かい分類ができていると考えられる．特に図中の S1，S2，S3，S4 に多くの電子メールが出力されており，これらのセルに着目すると，6 つのメーリングリスト中の 3 割以上の電子メールがマップ中の単一のセルに出力された．そのため，単一のセルに出力された電子メールの満たす条件に着目することで各セルの特徴を表現できると考えられる．以上の考察に基づいて電子メール分類ルールの生成を行う．

ルールの生成は生成された電子メールマップ中の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．以下に電子メール分類ルールの生成処理の流れを示す．（1）マップ中のあるセル (x,y) 座標に分類されたスレッドのリスト $Thread_{xy}$ を取得する．（2）$thread_{xyn}(thread_{xyn} \in Thread_{xy})$ の満たす条件 c を取得する．このとき inSubject(w) を抽出するため，電子メールのサブジェクトを形態素解析し，それぞれの形態素を w に代入した条件を用いる．また，inFrom(add)，inTo(add)，inCc(add) 及び inCommunity(add) について，対象とするスレッド中の電子メール全てがそれぞれの要素に対してメールアドレス add を含む場合，add にメールアドレスを代入して条件として用いる．（3）全メールスレッド中から条件 c を満たすスレッドの数を N(c) とし，c を満たす $Thread_{xy}$ 中のスレッドの数を $n_{xy}(c)$ とする．（4）N(c) が 3 以下の条件 c は電子メール分類ルール生成に用いないため，削除する．（5）各スレッド $thread_{xyn}$ が満たす条件を AND でつないだ組み合わせについて（3），（4）の処理を行う．（5）$n(c)/N(c) > \alpha$ 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．

4 実験

表 4 に図 1 に示したマップから生成された分類ルールを示す．R1. は電子メールの集合からメーリングリスト Core-Java-Technology 及びメーリングリスト java-dev-apple を単一のスマートメールボックスに分類する．R2. は電子メールの集合からメーリングリスト jsai の電子メールからサブジェクトに CFP と書かれた電子メールを単一のスマートメールボックスに分類する．本ルールの生成された理由としては，イベント情報に関するテンプレートの出現の有無が原因となったと考えられる．本ルールは，単純に人工知能学会のメーリングリストメールを分類するのではなく，さらに CFP（会議情報）に関する電子メールを分類する点で，有用である．R3. は電子メールの集合からメーリングリスト cocoa-dev 及び，メーリングリスト gnujava を単一のスマートメールボックスに分類する．R4. は電子メールの集合からメーリングリスト applescript users を単一のスマートメールボックスに分類する．Java に関するメーリングリストがまとめられるルールとして R1 及び R3 がある．これは R1 に分類された電子メールには電子メール中にプログラムソースが含まれていないものが多く，R3 には含まれているものが多かったためだと考えられる．

表 2: 生成された分類ルール

	生成されたルール
R1	inSubject(Core) & inSubject(Java) & inSubject(Technologies) & inSubject(Tech) & inSubject(Tips) & inFrom(SDN - Core Java Technologies Tech Tips) & inFrom(sunmail@hermes.sun.com) & in Community(SDN - Core Java Technologies Tech Tips) & inCommunity(sunmail@hermes.sun.com) | inTo(java-dev@lists.apple.com)
R2	inSubject([jsai-ann) & inSubject(CFP) & inFrom(admin@ai-gakkai.or.jp) & inTo(jsai-ann@iijnet.or.jp) & inCommunity(admin@ai-gakkai.or.jp)
R3	inCc(cocoa-dev@lists.apple.com) | inTo(java@gcc.gnu.org)
R4	inTo(applescript-users@lists.apple.com)

5 まとめ

本稿では従来の単語の出現頻度に基づく電子メールの特徴化に加えて，定型性に基づく特徴化を用いた自己組織化を行う，スマートメールボックスルール自動生成手法を提案した．近年，電子メールに SOM を用いた電子メールの分析を行う手法が提案されている．MailSOM[2] では，電子メールに出現する単語の出現頻度を特徴として用い，電子メールの分類マップ生成を行うシステムを提案している．しかし，Daniel らは電子メールの出現単語の頻度以外の特徴に着目していない，また，応用を行っていない．本研究では，電子メールの定型性に着目し，電子メールに SOM を適用した結果からスマートメールボックス生成に応用した．

参考文献

[1] 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)

[2] Daniel A. Keim, et al., ”MailSOM - Visual Exploration of Electonic Mail Archives Using Self-Organizing Maps”, In Proc. of International Conference on Email and Anti-Spam, 2005