投稿情報に基づくビジネスメールの話題分類

仲野 亘† 國分 智晴† 真鍋 俊彦† 坪井 創吾† 布目 光生†
†(株) 東芝 研究開発センター 知識メディアラボラトリー

1 はじめに

業務の電子化に伴い，プロジェクトやテーマといったコミュニティごとにメーリングリストを作成し，担当者間でコミュニケーション情報を共有することが行われている．メーリングリストを利用して業務を進めると，コミュニケーション情報が電子的に蓄積されるため，メールの中に保持された情報を知識として再利用することができる．例えば，過去の業務内容を思い出すことや，コミュニティに途中から参加した際にそれまでの業務内容を把握することなどが可能になる．しかし，企業内でやり取りされるメールの数は膨大なため，そのままの状態では蓄積されたメールを知識として活用することが非常に困難である．メールの整理のために一般的に用いられている手法はフォルダと仕分けを用いた手法である [1] が，設定のコストも高く，知識の再利用としての活用には繋がらない．

大量のメールを整理する手法として，受信したメールを内容に基づいて自動的に分類する手法が提案されている．例えば，Cselle らは一般的なメールをサブジェクトや送信者，宛先などの素性を利用して，逐次的に分類する手法である BuzzTrack を提案している [2]．ただし，ビジネス上のメーリングリストによるやり取りでは，宛先の情報が得られないことや，コミュニティに属するメンバ全体に宛てたメッセージがあるなどの特徴があり，一般的なメールに対する分類手法をそのまま適用するのは難しい．

本稿では，BuzzTrack のメール分類手法を参考に，ビジネスメーリングリストのメールを話題ごとに逐次的に分類する手法を提案する．また，提案手法を実業務上のコミュニティのデータに適用した結果について報告する．

2 メール分類手法

メーリングリストを介したやり取りでは，メールへの返信によりスレッド構造が構築される．返信は多くの場合，関係のあるメールであるという意図をもって行われるため，スレッドを最小単位としたメールの話題分類を行うこともできる．しかし，コミュニティによってはスレッド内で話題が切り替わる可能性があるため，提案手法では，主に個々のメールの内容の類似性を利用し，スレッド情報は副次的に利用する分類を行う．

また，進行中の業務に対してメールを整理する必要があるため，すでに活動が終了して投稿されなくなったコミュニティのメールを分類するのではなく，メールが投稿される度に逐次的に分類を行う．

2.1 分類アルゴリズム

提案手法のメール分類アルゴリズムは，BuzzTrack の手法に従う．まず，新規に受信したメールと，既存の全てのメールとの類似度を次節で述べる式に基づいて個別に計算する．これをメール類似度とする．次に，分類閾値 T を設定しておき，メール類似度の最大値と比較する．メール類似度の最大値が分類閾値 T よりも大きい値であった場合，新規メールをメール類似度の最大値が得られたメールが含まれる話題に分類する．メール類似度の最大値が分類閾値 T 以下の値であった場合は，新規メールは既存の話題には属さないとし，新しい話題のグループを作成して，その話題に分類する．

2.2 メール類似度

Cselle らの報告 [2] ではメールの話題分類に有効な素性としてメールのサブジェクト，本文，送信者と宛先の3点を挙げている．しかし，メーリングリストでは1節で述べたように，メールはメーリングリストのアドレス宛に投稿されるため，フォームから宛先の情報が得られず，また，メンバ全員宛に投稿されるメールも多い．そのため，送信者と宛先の情報をそのまま分類に用いるのは難しい．

ビジネスメールの特徴として，サブジェクトはメール内容を端的に伝える目的で書かれることが通常のメールよりも意識されることや，スレッド構造が広く利用されることがある．そこで，提案手法ではサブジェクトの類似度を主とし，そこにスレッド構造を考慮した補正を加えることで基本メール類似度を算出する．さらに，基本メール類似度に対し，実業務上のコミュニティのデータを分析することで得られた知見を元に2種類の補正を導入する．

基本メール類似度

サブジェクトの類似度 Sim_{sub} を以下のように算出する．まず，サブジェクトの文字列から，送信，返信時に自動挿入されるコミュニティ名などの定型部分を除去する．次に，残った文字列を形態素解析し，以下の式によりサブジェクトの類似度を算出する．

$$Sim_{sub}(m_i, m_j) = \frac{2|S_i \cap S_j|}{|S_i| + |S_j|}$$

ここで，S_i はメール m_i に上記の処理を適用した結果，名詞と判定された形態素の集合である．

また，同一スレッドに属するメール間の類似度にパラメタ Th 倍の補正を加える．サブジェクトの類似度 Sim_{sub} に Th を掛けたものを，基本メール類似度とする．

投稿間隔補正

閾値 T_d 日以上の投稿間隔があり，かつ異なるスレッドに属するメール間のメール類似度に，補正値 D を掛ける．この補正を投稿間隔補正とする．

Topic Detection and Clustering for Business Email based on Communication Patterns
†Wataru Nakano, Tomoharu Kokubu, Toshihiko Manabe, Sougo Tsuboi and Kosei Fume
†Corporate R&D Center, Toshiba Corporation

	コミュニティA	コミュニティB
データ採取期間	07/04/14-07/02	06/11/15-12/15
総メール数	248	162
投稿者数	9	9
スレッド数	61	57
正解話題数	54	36
正解話題中のメール数 (平均)	4.6	4.4

表 1: コミュニティ統計データ

	コミュニティA	コミュニティB
正解	- (54)	- (36)
従来手法	74.1 (40/54)	75.0 (27/36)
基本メール類似度	74.1 (40/54)	83.3 (30/36)
+投稿間隔補正	77.8 (42/54)	83.3 (30/36)
+人物補正	75.9 (41/54)	80.6 (29/36)
+両補正とも	79.6 (43/54)	80.6 (29/36)

表 2: コミュニティごとのメール分類の話題再現率

人物補正

閾値 T_p 通以上のメールが含まれる話題内のメールと新規メールとのメール類似度算出において，新規メールの送信者が話題内のメールの送信および本文から得られる宛名に登場していない場合，すなわち新しい人物が話題に割り込んだ場合，補正値 P を掛ける．この補正を人物補正とする．

3 実験

提案手法を実業務上のコミュニティデータに対して適用し，評価実験を行った．

3.1 実験概要

コミュニティA，Bの2つのコミュニティでやり取りされたメールを対象に分類を行った．このうち，コミュニティAは先に内容を分析し，提案手法の構築に利用したクローズドデータとして，コミュニティBはオープンデータとして用いた．

以下に各コミュニティの特徴について述べる．コミュニティAは離れた3拠点で作業を行う開発者の情報交換用に用いられた業務コミュニティである．システム開発の立ち上げから仕様検討，設計，実装に関する議論や連絡が主な内容であり，緊急の要件以外のコミュニケーションは基本的にコミュニティ上で行われた．

コミュニティBはある拠点で作業を行う研究者の情報交換用に用いられた業務コミュニティである．主にマネージャから作業の指示や報告の要求があり，各担当者がそれに答えるという形式が採られている．

これらの2つのコミュニティに対し，話題ごとのメール分類を手動で行って，正解データとして整備した．コミュニティAは著者の1名が分類した結果を，コミュニティBは著者のうち4名が分類を行った結果を統合した．表1は各コミュニティの統計量と，作成した正解データの統計量をまとめた表である．

この正解データを用いて提案手法の評価を行う．なお，分類結果の話題のうち，含まれるメールが正解の話題と完全に一致する話題の数を正解話題数で割った値を話題再現率とし，分類精度の指標として用いる．

3.2 実験結果

表2はコミュニティごとの分類結果における話題再現率を示した表である．なお，分類閾値は $T = 0.7$ とし，その他のパラメタは $Th = 1.5$，$T_d = 10$[日]，$T_p = 8$[通]，$D = 0.2$，$P = 0.2$ とした．パラメタはコミュニティAに対して補正2種を両方加えた提案手法を適用した際に，最も高い話題再現率を得た値を設定している．また，BuzzTrackのメール分類手法 [2] のうち，コミュニティAに対して行った予備実験において最も話題再現率の高かった，サブジェクトの類似度による分類手法の適用結果を従来手法として比較した．

コミュニティAに対しては，基本メール類似度では従来手法から話題再現率に変化はないが，2種類の補正により話題再現率が向上し，2種類を組み合わせると上昇量も合計される．サブジェクトとスレッドの情報を用いた基本メール類似度，および2種類の補正がメールの話題分類に効果的に働いていることがわかる．

コミュニティBに対しては，基本メール類似度により話題再現率は向上するが，投稿間隔補正は効果がなく，人物補正は僅かではあるが再現率を下げてしまっている．コミュニティBで1件発生した失敗について分析したところ，話題に割り込んだ人物のコミュニティにおける役割の違いがあった．コミュニティAでは，プロジェクトのオブザーバやリーダなどが割り込み，話題を転換・修正させる事例であったのに対し，コミュニティBでは，マネージャと担当者間のやり取りに主担当者がフォローのために割り込むという事例であった．前者では話題が切り替わり，後者では話題が継続されたため，分類に失敗していた．コミュニティの性質や割り込む人物の役割が影響することから，適切にフィードバックを与えることにより，コミュニティごとに人物補正の効果を調節できる可能性がある．

4 おわりに

本稿では，蓄積されたメールからユーザの作業目的に対応した知識を素早く適切に取得することを目的とし，ビジネスメーリングリストのメールを，メーリングリスト特有の問題に対応して話題分類する手法を提案した．また，提案手法を実業務上のコミュニティに適用し，評価結果を通してメール分類の効果とビジネスメールによるコミュニケーションの特性について報告した．分類実験の結果，クローズドデータに対して最大79.6%，オープンデータに対して最大83.3%の話題再現率が得られ，ビジネスメーリングリストに対する話題分類が従来手法より高い精度で可能なことを確認した．

今後は，対象データの規模を拡大し，分類手法の改良やコミュニティごとの調整を行う．また，分類したメールを元に，ユーザの状況に応じ，適切なメールを話題ごとに自動で提示するなど，メール分類手法の活用法についても検討する．

参考文献

[1] D. Fisher, A. J. Brush, E. Gleave, and M. A. Smith. Revisiting Whittaker & Sidner's "Email Overload" Ten Years Later. In *Proc. CSCW'06*, pp. 309–312, 2006.

[2] G. Cselle, K. Albrecht, and R. Wattenhofer. BuzzTrack: Topic Detection and Tracking in Email. In *Proc. IUI'07*, pp. 190–197, 2007.