マイクロブログサービスの返信行動に着目した投稿及びユーザの分類

黒澤　義明　　竹澤　寿幸
広島市立大学大学院　情報科学研究科
{kurosawa, takezawa}@ls.info.hiroshima-cu.ac.jp

1.　はじめに

近年, マイクロブログサービスが爆発的に普及している. マイクロブログサービスが関心を呼ぶ理由として, 田中ら(2010)が指摘するように, blog のような使い方, SNS (Social Network Service)のような様々な使い方ができること, そして学問的にも複数の位置付けが可能であり, その切り口が多様になることも挙げられる.

本研究は, マイクロブログに少なくとも 2 つの大きな位置付けがあると考える. 1 つは情報発信を目的とした, ブログの一種との位置づけ, そしてもう 1 つは, SNS の一種としての位置づけである. 前者の側面を利用して様々なメディアや企業が参入している. 後者は, 何らかのコミュニティ内で人とのつながりを求めるユーザ～おそらく一般ユーザ～が主に重視している側面と考えられる.

本研究は今回, 一般ユーザに対して, より有効と考えられる SNS 機能としてのマイクロブログサービスに注目する. 特に, ユーザが行う返信行動に着目し, その元投稿及び返信内容から投稿内容ベクトルを生成した上で, ユーザのクラスタリングを行う. そして, 獲得クラスタと現実生活に存在するコミュニティとの比較を行うこととする.

2.　twitter

本研究は, マイクロブログサービスの中でも, twitter に焦点を当てる.　twitter が他の SNS と異なる点は, 相互に許可が要らない点, すなわち, コミュニティへの加入許可が要らない点である.

2.1.　フォローとフォロワー

twitter では, あるユーザに対しフォローと呼ばれる手続きを行うことにより, そのユーザのフォロワーとなり, そのユーザの投稿が閲覧可能となる. 相互にフォロー行為を行うことは閲覧のための必須条件ではない. このため, SNS のようにはっきりとした境界を持つコミュニティは形成されず, 緩やかなコミュニティのみが形成される.

2.2.　返信行動によるコミュニティの推定

フォローという行為を行う理由として,「友達だから」「有名人だから」などの様々な理由が考えられる. 有名人は一般ユーザに対してあまりフォローしないことを考えると, 相互フォローを行う関係に着目することにより, 現実世界のコミュニティを発見できる（畑本ら 2010）. しかし, 実際にはフォローしただけで互いに全く交流がないユーザもおり, 相互フォロワーという指標が最適かわからない（岩本ら 2009）. そこで, 返信行動に着目する.

前述の通り, あるユーザをフォローさえすれば, そのユーザの投稿は読める. このため, コミュニティに属していなくとも返信は可能である. しかしながら, こうしたユーザの返信回数は多くないことが予想される. したがって, 返信行動が多数ある複数のユーザの投稿調査により, 所属コミュニティが検出できると考えられる.

2.3.　ユーザの興味の推定

同じコミュニティに属しているとは言え, 全ての投稿に対し, 返信を行うことはない. おそらく, ユーザは興味のある話題のみに返信すると考えられる. 岩木ら(2009)もこうした仮説に立ち, 返信中の特徴語に着目することにより, ユーザと似たブロガーを見つけ, 有用な記事の発見支援を試みている. ユーザのクラスタリングを行う本研究と目的が異なるとは言え, 基本的な考え方は似ている.

ただし, 彼らの指標は, 特徴語の類似度計算(similarity)と, 返信回数(connection)がそれぞれ計算されているため, 本研究はより直接的にユーザ間の興味が共通化されるような手続きを行う. 返信のカテゴリ化と, 投稿-返信の双方に対する興味の共通化手続きである.

3.　本研究の提案内容

次に本研究の提案内容を述べる.

3.1.　返信対のカテゴリ共有

本研究はユーザのクラスタリングを有効に行うため, ユーザ間の興味の共通化を行う. 本研究における仮定を示す.

① 元投稿と返信の間で両ユーザの興味は同一
② 興味は, カテゴリによって記述可能

例えば, 以下の投稿対を考える.

T「ネコいた！ ネコってかわいいよね^^」
R「イヌに一票(^^)／」

このとき, T と R の投稿を全く異なる対象（ネコとイヌ）についてなされたという解釈ももちろん可能である. しかし, 本研究では①の成立を仮定する. さらに, 元投稿と返信の間には共通の話題（カテゴリ）が存在すると考える（図 1）. 図中,『ネコ』は「ペット」「ネコ科」「家畜」というカテゴリに,『イヌ』は「ペット」「家畜」「モデル生物」というカテゴリに属することを示す.

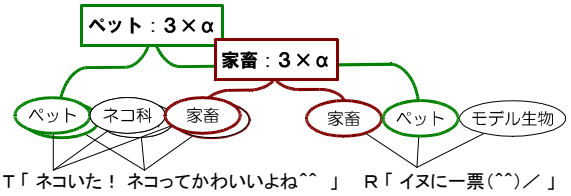

図 1　返信対の例

この返信対では，カテゴリ「ペット」が 3（ただし，重みαが付く），「家畜」も 3（重みα付），「ネコ科」が 1（重みなし），「モデル生物」が 1（重みなし）という値となる．したがって，①②の仮定に基づき，ユーザ T，ユーザ R の興味内容を示す投稿ベクトルはともに，(ペット, 家畜, ネコ科，モデル生物）＝（3α，3α，1，1）となる．この手続きにより，繰り返し同カテゴリについて投稿-返信を繰り返すユーザは，互いに似た投稿ベクトルを持つに至る．

なお，本研究は Wikipedia 辞書 [1]を用いて，カテゴリの特定を行う．青島ら(2010)が行ったように日本 WordNet 利用することもできよう．ただ，Wikipedia 辞書の方が登録語数も多く，最新の項目が追加されるため，語が未登録であるために，ユーザの分類が行われないことを防ぐ効果があると考え，同辞書を用いることとした．

3.2.　関連研究について

青島ら(2010)も述べているように，投稿は短く，単一投稿だけで頻度等の特徴量記述を行うことは難しい．本研究は，単一投稿で不足する情報を，返信行動に着目し，カテゴリ化を行うことで補った研究であると言える．

このような不足情報を補うことを目的とした研究として，本研究同様，返信を使った研究（岩木ら 2009），制約情報をもとに投稿の類似度を使った研究（青島ら 2010）リンクを持つ投稿に着目した研究（ex. 吉田ら 2009），「お気に入り」に着目した研究（真野ら 2010），ユーザが分類した「リスト」に着目した研究(榊ら 2009)，RT に着目する方法（向井ら 2011）等が挙げられる．この他にも，ハッシュタグに注目することも考えられる．

これらの有効性の比較のため，投稿の傾向についての事前調査結果を表に示す（表 1）．表中 HCU とは，本学関係者と思われるユーザ 298 名のことである．一般的に，投稿の 3 割程度は返信であり，他の指標に着目するよりも，少なくとも数の上では有効となることが予想される．

4.　自己組織化マップ SOM による視覚化

本研究では，投稿ベクトルの視覚化に Kohonen(2001)による自己組織化マップ(Self-Organizing Map, SOM)を使用する．SOM は，多次元ベクトルデータをその特徴を

[1] page, redirect, categorylinks の各 mysql データを用いた．redirect の使用により，表記の揺れにも一部対処可能である．

表 1　ツイートの傾向

	Streaming API ～ sample ～ 20101201-20110119		HCU 20101201-20101226	
	%	tweet	%	tweet
返信	36.2	2,986,408	27.2	82,183
リンク	9.7	803,022	5.0	15,122
ハッシュタグ	5.6	464,231	6.8	20,561
公式 RT	3.4	282,565	0.9	2,665
非公式 RT	4.5	368,357	1.8	5,539
非公式 QT	0.4	32,663	0.3	951
総ツイート		8,247,607		301,591

残したまま，2 次元マップに写像する．特に非線形のデータに対し有効であり，Kurosawa et al.(2010)による擬情語の分類等，自然言語処理での有効性が確かめられている．

4.1.　自己組織化マップのアルゴリズム

SOM は二層からなる神経回路網モデルである．入力層への入力により，競合層の特定の領域が反応するような，教師なし学習を行う．

入力層への n 次元の入力ベクトル x は，$x=\{x_1, x_2, \cdots, x_n\}$ と表現する．また，競合層にはノードと呼ばれるユニットがあり，全ノードから，入力層との間に参照ベクトル m と呼ばれるリンクが行われる（図 2）．

ここで，次式を満たす勝者ノード c の発見を試みる．次式は入力ベクトルに最も類似した参照ベクトルを持つノードを見つける操作と考えられる．

$$\forall i,\quad \|x-m_c\| \le \|x-m_i\|$$

勝者ノードの発見に続いて，近傍 $h_{ci}(t)$を決める．本研究では時間 t とともに減少するガウス関数を用いた．この近傍内では，複数の参照ベクトルを入力ベクトルに近づける操作を行う．つまり，時間が経つにつれ，近隣のノードの類似性が増し，隣接ノード間距離が近づく（図 4 の右中央部の変化）．以下，時間 t を用いた更新式を示す．

$\forall i \in N_c(t)$ を満たすとき，

$$m_i(t+1)=m_i(t)+h_{ci}(t)(x(t)-m_i(t))$$

それ以外のとき，　$m_i(t+1)=m_i(t)$

以上の勝者ノード発見，近傍更新を繰り返すことにより，学習を行う．これが SOM のアルゴリズムである．

5.　実験と考察

5.1.　言語データ

本研究で用いる実データの収集・加工手続きを示す．

①　応答対の取得

先述の 298 人の投稿（2010/12/01～2010/12/26），約 30 万件のデータ，8 万件の返信の中から，投稿-返信ともに本学関係者である 51,220 応答対を得た．

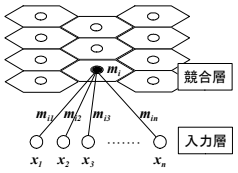

図 2　SOM の基本概念

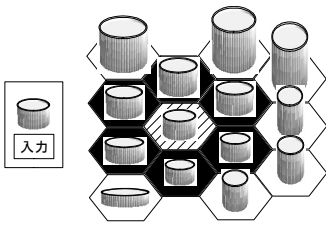

図 3　勝者ノード，近傍

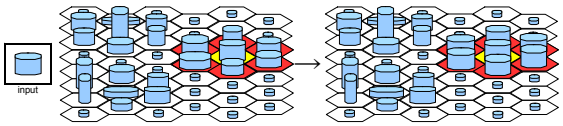

図 4　参照ベクトルの更新

② ユーザ毎カテゴリ出現率算出

Wikipedia 辞書エントリを追加した MeCab により形態素解析を行った後，代名詞等を除く名詞に対し，Wikipedia 辞書による投稿内容のカテゴリ化を行った．2 階層上まで探索した結果，上記の投稿が 16,329 カテゴリにより表現された．また，応答対のうち 16,307 応答対に，図 1 の例に挙げたような共通のカテゴリが存在した．なお，今回の実験ではこの共通カテゴリが分類に寄与するよう，共通カテゴリを持つ投稿に重み付け（総カテゴリ数×10）を行った

その上で，ユーザの興味をまとめるため，ユーザ毎カテゴリ毎に出現頻度を求め，さらにユーザのツイート数の影響を減じるため出現率に換算した．

$$\text{あるユーザのカテゴリ } C_i \text{の出現率} = C_i \Big/ \sum_{i=1}^{n} C_i$$

③ pLSA による次元圧縮

カテゴリ数の増加により，計算時間増加の問題が生じるため Hoffman(1999)による pLSA(probabilistic Latent Semantic Analysis)の工藤の実装を用い，次元縮約を行い 150 に圧縮した．なお，温度パラメータ β=1.0（厳密な EM の実行）を採用した．

5.2.　実験手続き

2 章で説明した手続きにより，som_pak を使用した 2 段階の分類学習を行った．予備実験により決定された初期学習率係数 α，初期近傍半径 r のパラメータを以下に示す．

マップサイズ：64 ノード×48 ノード

1st: 学習回数 1,000,000，α=0.05，r=80

2nd: 学習回数 10,000,000，α=0.01，r=40

5.3.　実験結果と考察

実験結果を図 5 に示す．隣接ノード間距離の最大値と最小値を元に，距離が 0-1 になるよう変換し，明度で表現した図である．また，図 6 にマップの見方を示す．

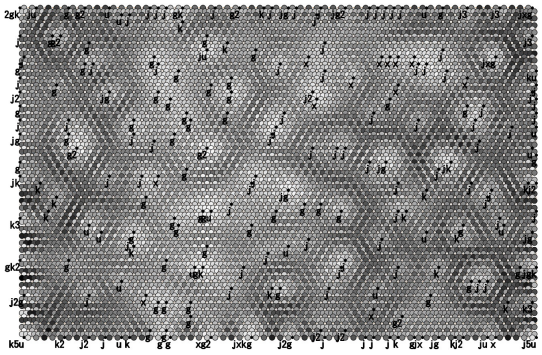

図 5　SOM によるユーザ分類結果

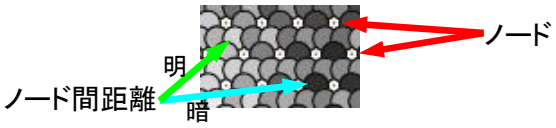

図 6　マップの見方

赤線の先に示した小丸がノードを示す．また，図 6 中，扇のような形状により，ノード間距離を示す．白（緑線）が近く，黒が（水色線）遠いことを示しており，グラデーションにより表現した．暗い輪郭を持ち，かつ明るい内部を持つ領域は，外側とは異なる特徴を持っていると考えられる．つまり，コミュニティがあると考えられる．

なお，該当ノードにユーザが存在する場合には，*jxgku* の所属学部と人数が付与される．ここで，*jgk* は本学の 3 学部を指す．そして，*u* は所属不明を表す．なお，*x* は著者が所属する研究室（学部 *j* に属する）である．今回のデータについて，ユーザの所属とその人数を示す（表 2）．

表 2　所属の内訳

学部				不明
j		*g*	*k*	*u*
x 研究室	その他			
19	129	89	40	21

5.3.1.　周辺部について

マップ左下に「*k5u*」，右下に「*j5u*」等，一部のノードに同一学部ユーザが配置されている．したがって，一部の特徴的なユーザをクラスタリングできたことを示す．

しかし，周辺部の多くのユーザは単独ノードで存在しており，領域としてはまとまっていないように見える．この原因としては，pLSA による 150 次元への次元数が多すぎてクラスタにはならなかったこと，あるいは共通カテゴリに対する重み付けにより，共通カテゴリを持たない投稿対に効果が及ばなかった結果と考えられる．

また，周辺部のユーザは返信回数が少なく，特定のカテゴリだけが強調されたユーザでもある．この結果は本手法の出現率を求める方法にあると考えられる．ユーザの投稿数で補正を行う必要について，さらに検討したい．

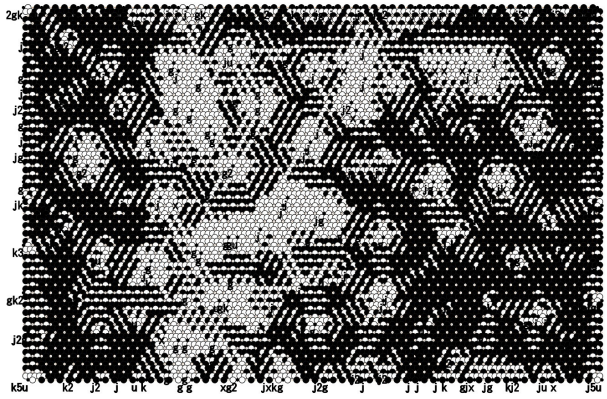

図 7　強調したユーザ分類結果

5.3.2.　中央部について

中央部は多くの返信行動を行ったユーザであり，多くのカテゴリを利用したユーザでもある．このため，領域が明確でないところもあり，図に強調化を施す．今回は，0.33 以上の距離を持つノードを黒，それ以外を白で着色した（図 7）．また，図の一部を拡大して示す（図 8）．

黄色の部分には，g が 7 人，j が 1 人である．また，その近辺に比較的多くの g が配置されていることから考えて，特定の学部のユーザが収集できたと考えられる．水色の部分についても，7 人の学部 j が配置されており,こちらも特定の学部のユーザが収集できたと考えられる．

5.3.3.　学部以外の分類について

図 8 の赤色の領域には 4 人の j と 4 人の x が含まれている．両者は学部 j であるため，全体としては正しく分類されていることがわかる．ただし，x が入るであろうサブクラスタの分類はできていないことになる．

実際に所属するクラスタという観点からは，畑本ら(2010)で行ったような，フォロー・フォロワー関係を用いた方がより適切な実験結果が得られる可能性がある．一方，興味に応じた返信を基準とするクラスタという意味においては，本手法に基づく赤色の領域分けが正しいはずである．ただ，実際に所属するクラスタと，興味に基づいたクラスタの使い分けは今後の課題である．

5.3.4.　問題点

Wikipdeia の辞書によるカテゴリ化が不適切となる点が問題である．例えば，「なんだろう」という表現が特定のキャラクターになる等である．今回は機械的に登録したことによる．今後，平仮名回避等の処理が必要となる．

6.　おわりに

本研究は twitter の返信行動に着目し，投稿ベクトルのカテゴリ共通化・クラスタリングを行った．クラスタリング結果と現実の所属とを比較し，本機能の有効性を確認した．クラスタの特徴から，ユーザの特徴を抽出する，あるいは提案のために活用する等は今後の課題である．

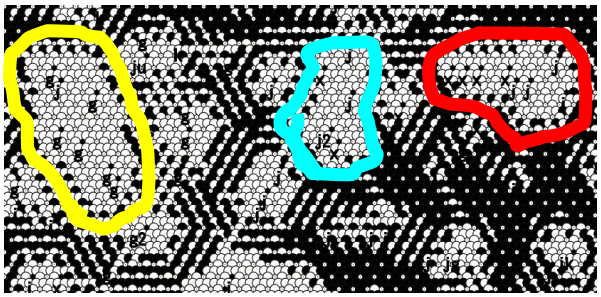

図 8　分類結果（一部拡大）

参考文献

青島傳隼，福田直樹，横山昌平，石川博 (2010)．“マイクロブログを対象とした制約付きクラスタリングの実現.” DEIM2010.

畑本典宣，黒澤義明，目良和也，竹澤寿幸 (2011)．”マイクロブログにおけるユーザのクラスタリングとそのクラスタの特徴語抽出．” 言語処理学会第 17 回年次大会.

Hofmann, T. (1999). ”Probabilistic Latent Semantic Indexing.” in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval”, pp.50-57.

岩木祐輔，アダム ヤトフト，田中克己 (2009)．“マイクロブログにおける有用な記事の発見支援.” DEIM2009.

Kohonen, T. (2001). “Self-Organizing Map, 3rd Edition.” 徳高平蔵，岸田悟，藤村喜久郎訳 (2005) “自己組織化マップ.” シュプリンガー・ジャパン.

工藤拓. “PLSI”, http://chasen.org/~taku/software/plsi/

工藤拓．“形態素解析器 MeCab.”, http://chasen.org/~taku/software/mecab/.

Kurosawa, Y., Mera, K., and Takezawa, T. (2010). “Psychomime Classification and Visualization Using a Self-Organizing Map for Implementing Emotional Spoken Dialog System.” In Spoken Dialogue Systems Technology and Design, Wolfgang Minker, W., Lee, G. G., Nakamura, S., and Mariani, J. (eds), pp.107-134, Springer.

眞野裕也，青山俊弘 (2010)．“ミニブログユーザの記事嗜好を用いたクラスタ発見.” 日本高専学会誌，15(3), pp.43-46.

向井友宏，黒澤義明，目良和也，竹澤寿幸 (2011)．”マイクロブログの分析に基づくユーザの嗜好とタイミングを考慮した情報推薦手法の提案．” 言語処理学会第 17 回年次大会.

som_pak, “som_pak.” http://www.cis.hut.fi/research/som_pak/

榊剛史，松尾豊 (2010)．“ソーシャルブックマークとしての Twitter リスト機能の応用.” The 24th Annual Conference of the Japanese Society for Artifcial Intelligence.

田中淳史，田島敬史. (2010)．“twitter のツイートに関する分類手法の提案.” DEIM2010.

Twitter, “Twitter.” http://twitter.com/

Wikipedia, “Wikipedia 日本語版.” http://ja.wikipedia.org/

Wikipedia, “Wikipedia:データベースダウンロード.” http://download.wikimedia.org/jawiki/

吉本和紀，鈴木優，吉川正俊 (2010)．“マイクロブログにおける他者への影響を考慮した投稿者の重要度推定手法.” DEIM2010.

吉田光男，乾孝司，山本幹雄 (2010)．“リンクを含むつぶやきに着目した Twitter の分析.” DEIM2010.