Twitterから有益な日常情報を発見するための特徴語による地域ユーザの検索

田原　琢士†　　馬　　強††

† 京都大学工学部情報学科　〒 606–8501 京都府京都市左京区吉田本町 36-1
†† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町 36-1
E-mail: †tahara@db.soc.i.kyoto-u.ac.jp, ††qiang@i.kyoto-u.ac.jp

あらまし　Twitter などのマイクロブログで発信されている地域・日常生活に役立つ “その場” の情報を発見するには，地域情報を発信するユーザの検索が重要な課題である．そこで，本研究では出現頻度，発信ユーザ数と発信時間，およびその他の地域との差異を考慮して対象地域の特徴語を抽出し，それに基づく地域ユーザの検索手法を提案する．
キーワード　マイクロブログ，Twitter，地域情報，地域ユーザ検索

1. はじめに

現在，マイクロブログが非常に普及している．その代表である Twitter(注1)では 140 字以下の Tweet と呼ばれる短文を投稿でき，ユーザは日常における体験をその場で発信することができる．このような情報の中には，ユーザの生活地域における有益な情報，例えば生活地域内でのイベントや，交通の情報，お店の混雑状況といった情報が含まれている．これらの情報はその時点，その場所で役立つ場合が多く，リアルタイムに知ることができれば地域ユーザの日常生活に有用であると考えられる．

このような日常生活における有益な情報を Twitter から獲得するには，Tweet の発信時間と場所を把握する必要がある．Twitter では発信時間は明記されているが，場所に関する情報が不十分である．Tweet がどの場所で発信されたかが事細かに記載されることは稀である．また，Twitter の機能として位置情報のタグ（ジオタグ）を付けることは可能であるものの，その利用率は 2012 年の時点でわずか 0.77%(注2)に過ぎない．

地震や有名人の動向など多くのユーザが関心もつホットな情報であれば，榊らの研究 [1,2] で示されているように，必要なジオタグつき Tweet を集めることができる．しかし生活地域での日常生活における有益な情報に関心を持つのはその生活地域内のユーザであるため，日常生活における有益な情報を示す充分なジオタグつき Tweet を得るのは難しいと考えられる．

ここでの生活地域というのは小さな範囲である．例えば，現在隣の市でドリンクの無料配布が行われていると言う情報を知っても，ユーザはわざわざもらいに行かない．しかし自分が歩いて行ける場所で配布をしているならば，ユーザはもらいに行くであろう．このように，生活地域での日常生活における有益な情報というのは，興味を持つユーザの活動範囲が小さな地域に限定されるのである．

そこで，ジオタグ付きのツイートから地域・日常生活に関する情報を発見するのでなく，その生活地域内の Twitter ユーザを検索して．それらのユーザの発言を対象に “その場” の地域情報を獲得することが有効であると考えられる．小さな範囲の地域ユーザの検索のために，本論文では特定地域のユーザとその周辺の他の地域のユーザの Tweet から，語の出現頻度，語を発信するユーザ数と語が発信された日数，特定地域とその他の地域の差異を考慮して語の地域度を算出することで対象地域の特徴語を抽出し，その地域特徴語を用いて地域ユーザの可能性のあるユーザを発見したのち，地域特徴語の地域度を成分にとる地域ベクトルを用いてーザ分類を行い，地域ユーザを検索する手法を提案する．

2. 関連研究

Twitter からイベント情報を発見する研究としては，[1,2] 以外にも，路線名を含んだ Tweet から SVM による分類で鉄道のトラブル状況を抽出する土屋らの研究 [3] や，地名を含んだ Tweet から単位時間当たりの呟き数を考慮することでイベントなどを検出する松村らの研究 [4]，foursuqare(注3)を経由した Tweet と，それらに記載された地名を含んだ Tweet から共起する語を抽出することで，その場所でのイベントを発見する渡辺らの研究 [5]，一定時間内に成されたジオタグ付き Tweet からユーザのクラスタを作成し，クラスタ内での共起語を抽出してイベントを発見する杉谷らの研究 [6] などが挙げられる．

[3] の研究では東京メトロの 9 路線についてトラブル状況を検出しており，鉄道事故は路線をその時点で利用している多くの乗客にとって関心のあるイベントのため，路線名を含んだ Tweet だけからでもある程度有効な抽出ができている．[4] の研究では停電などの規模の大きなイベントを検出している．これらの研究の手法で検出できるのは場所の名前を含んでいる Tweet が示すイベントだけであり，そのほとんどが広い範囲の人々が関心を持つ規模の大きなイベントである．しかしながら我々が求める日常情報は関心のある人々が狭い範囲に限定され

(注1)：https://twitter.com
(注2)：http://semiocast.com
(注3)：https://foursquare.com/

るもので，この情報を示す Tweet に場所名が含まれることは稀であると考えられる．

[5,6] の研究では比較的小規模のイベントを検出している．[5] は foursquare に登録された場所のうち，有名な場所や駅名など地域によって示す場所が曖昧にならないものについてのみイベントを検出しているため，我々の求める日常生活の範囲での日常情報を検出しようとすると検出数が極僅かになってしまう．[6] の研究についても，ジオタグ付き Tweet が少なさのために，狭い範囲での日常情報の検出数は少なくなってしまうと考えられる．このため，まず狭い範囲の Twitter ユーザを発見し，それらのユーザの Tweet から地域情報を取得するのが有効であると考えられる．

Twitter ユーザの属性情報として生活地域を推定する研究としては，ツイート中の単語と地域の相関をもとにした確率モデルとユーザの位置推定を調整するための格子ベースの近隣平滑モデルを用いた Cheng らの研究 [7] や，LDA(Latent Dirichlet Allocation) を用いて Tweet のトピックを作成してユーザの生活に関わる地域を推定する堂前らの研究 [8] などがある．前者はアメリカの市レベルでの位置推定を行い，100 マイルの誤差で 51%のユーザの位置を特定している．後者は日本の都道府県レベルで位置推定を行い、精度 0.59，再現率 0.54 であった．これらに対して，本研究ではより小さい生活地域のユーザの検索を行う手法を提案する．

地域特徴語を用いてユーザの居住地を推定する研究として西村らの研究 [9] が挙げられる．西村らは地域の Tweet を一つの文書と見た TF-IDF 法に基づいて地域特徴語を抽出し，それを素性とした SVM による多クラス分類を用いて 47 都道府県を単位として居住地推定を行い，F 値 0.34 を達成している．この手法に対し，本研究では地域の Tweet がその地域の複数のユーザが多様な時間に呟いた Tweet であることを考慮し，語を呟いた地域ユーザの数と呟かれた時間も考慮して地域特徴語を検出する．

また長谷川らの研究 [10] では，TF-IDF 法に基づいて，まず特定の地名と共起度の高い語から地名に対する特徴語辞書を作成し，類似する複数の辞書を時間的・空間的に連続したものだと捉えて統合，地域の特徴語辞書とする．この操作により，長谷川らは地域特徴語が被覆する地域・時間の範囲をより詳細に求めることを試みている．この研究では辞書作成に用いる地名をあらかじめ与えて Tweet を収集するが，本研究ではそういったものは与えず，特定地域とその周辺の地域のユーザの Tweet を用いて地域特徴語を抽出する．

3. 地域ユーザの検索手法

特定地域の情報を日常的に発信するユーザは，その地域特有の場所・日常的な話題などを示す語を多く発信していると考えられる．そこで，本手法ではそういった地域特徴語を Tweet から抽出し，地域特徴語を多く呟くユーザをその地域の日常情報を発信するユーザとする．

3.1 地域特徴語の抽出

本研究では，特定地域の特徴語はその周辺の他の地域に比べて出現頻度が高い，もしくはその地域のユーザのみが多く発信する語とする．地域特徴語抽出の手順を以下に示す．

（1） 各ユーザの Tweet から既存の形態素解析ツールを用いて語を切り出す

（2） 各語の出現頻度，発信ユーザ数，発信日数，他の地域での出現及び出現頻度を考慮して地域度を計算する

（3） 地域度の高いものを地域特徴語とする

本節では 3.1.1 で (1) について，3.1.2 で (2) について述べる．

3.1.1 Tweet からの語の切り出し

Tweet からの語の切り出しには既存の形態素解析ツール（MeCab(注4),etc.）を使い，未知語と名詞のみを切り出す．ただし Li らの研究 [11] で述べられているように，Tweet には１４０字の制限があるために文法的な誤りや正式でない省略がみられる．このため，形態素解析ツールが未知語や名詞として推定するもののみを切り出してしまうと，本来意図されている語を切り出せない場合がある．そこで本手法では，形態素解析ツールが推定した連続する二つ以上の名詞または未知語を一つのまとまりと扱い，推定された単語を基本単位とする N グラム法 ($1 \leqq N \leqq 7$) を用いて連続する 1 つ以上の名詞または未知語を全て一つの名詞として切り出す．例えば MeCab は「熊野寮祭」を「熊野」と「寮」と「祭」の連続した三つの名詞として推定するが，本研究では「熊野」,「寮」,「祭」,「熊野寮」,「寮祭」,「熊野寮祭」の六つの語として切り出す．

また，上記の処理の際に出る 1 文字だけの平仮名やカタカナ，アルファベット，記号などは地域特徴語にはならないと考えられるので処理の際に取り除く．また，Twitter 上で他人の Tweet を拡散する「リツイート」を示す文字列「RT」や他人への返信を表す「@“ユーザ名”」，URL を表す文字列，顔文字，Tweet の話題を示すために用いられるハッシュタグを表す「#“文字列”」などの地域特徴語に当たらない語もできる限り取り除く．

3.1.2 地域度の計算

特定地域の特徴語は，出現頻度が周辺の他の地域の Tweet 内に比べて高く，その地域の Tweet 内にしか現れない語であると考えられる．しかしながら地域の Tweet というのは一人の人物が一度に書いた文書ではなく，実際には様々なユーザが異なる時間に発信した Tweet の集まりである．このため，各地域の Tweet を一つの文書として特定地域の Tweet での語の出現頻度と周辺の他の地域で語が出現するかどうかだけを考慮するのではなく，語を発信する特定地域のユーザ数及び特定地域で語が発信された日数も考慮すべきである．そこで，以下に示すような指標を考慮して地域度を計算する．

1. 語の出現頻度

前述したように各地域の Tweet は一つの文書ではなく，膨大な数の Tweet の集まりである．このため，地域での語の出現頻度を地域の Tweet 内での語の出現回数としてしまうのは問題がある．例として一個の Tweet に語が十個含まれている場合と，十個の Tweet に語が一個ずつ含まれている場合を考える．全体としてどちらも語の出現回数は十回であるが，ユーザ

(注4)：https://code.google.com/p/mecab/

が Tweet で語を発信した回数は前者は一回，後者は十回となる．従って，Twitter においては語の全体での出現回数ではなく語の Tweet による発信回数を重視すべきである．そこで本研究では地域 C_j での語 t_i の出現頻度 $tf(t_i, C_j)$ を以下のように定める．

$tf(t_i, C_j) = (C_j$のユーザが語 t_iを含む `Tweet` を呟いた回数)

2. 周辺の他の地域を考慮した語の出現頻度

1 で述べた出現頻度が高い語が地域の特徴語になるとは限らない．単純に出現頻度が高くなる語というのは，例えば ‘今日’ や ‘時間’ などといった地域に関係なく一般的によく呟かれる語である．我々が求めている特徴語はあくまで ‘周辺の他の地域に比べて’ 出現頻度が高い語である．この相対的な出現頻度を式 (1) で計算する．

$$rtf(t_i, C_0) = \frac{tf(t_i, C_0)}{\frac{1}{|C|}\sum_j tf(t_i, C_j)} \tag{1}$$

ここで C_j は各地域，C_0 は対象としている特定地域，$|C|$ は地域の総数をそれぞれ表す．式 (1) では語 t_i の地域 C_0 での出現頻度を全地域の語 t_i の出現頻度の平均で割っており，どの地域においても出現頻度が同程度の語，または地域 C_0 で出現頻度が他の地域より低い語は値が小さくなり，地域 C_0 で出現頻度が他の地域より高い語は値が大きくなる．

3. 周辺の他の地域での語の出現の考慮

特定地域の特徴語はその地域でしか発信されず，一般的な語はどの地域でも発信されると考えられる．そこで語が周辺の他の地域で出現しているかを考慮するために式 (2) を計算する．

$$icf(t_i) = \frac{|C|}{cf(t_i)} \tag{2}$$

ここで $cf(t_i)$ は語 t_i を含む Tweet を呟いたユーザが存在する地域の数である．式 (2) では総地域数 $|C|$ を $cf(t_i)$ で割っており，語 t_i が出現する地域の数が少ないほど値が大きくなる．

4. 発信ユーザ数

2，3 で述べた $rtf(t_i, C_0)$，$icf(t_i)$ の値が大きいほど語 t_i は特定地域 C_0 に特有の語である可能性が高い．しかしながら地域の特徴語は ‘その地域の中では’ 一般的な語であると考えられ，特定地域に特有の語であっても，その地域のごく一部のユーザだけが発信する語はその地域の特徴語ではないと考えられる．そこで，語を発信するその地域のユーザ数を考慮するために式 (3) を計算する．

$$uc(t_i, C_0) = \frac{u(t_i, C_0)}{|U_{C_0}|} \tag{3}$$

ここで $u(t_i, C_0)$ は地域 C_0 のユーザのうち語 t_i の含まれる Tweet を呟いたユーザの数，$|U_{C_0}|$ は地域 C_0 のユーザの総数をそれぞれ表す．したがって地域 C_0 において一部のユーザだけが呟く語は式 (3) の値は小さくなる．

5. 発信された日数

2，3，4 で述べた $rtf(t_i, C_0)$，$icf(t_i)$，$uc(t_i, C_0)$ の値が大きい語は，その地域特有の語であって多くのユーザに発信されていると考えられる．しかしそういった語の中で，短期間の間にだけ発信された語は突発的な事件やイベントなどを表す語である場合が多い．これらを除くために語が発信された時間 (日数) を考慮する式 (4) を計算する．

$$dc(t_i, C_0) = \frac{d(t_i, C_0)}{|D|} \tag{4}$$

ここで $dc(t_i, C_0)$ は地域 C_0 内のユーザが語 t_i を含む Tweet を呟いた日数，$|D|$ は総日数を表す．よって短期間に地域 C_0 のユーザが呟いた語は式 (4) の値が小さくなる．

本研究では以上で挙げた式 (1),(2),(3),(4) を用いて語 t_i の特定地域 C_0 の地域度を式 (5) で計算する．

$$loc(t_i, C_0) = rtf(t_i, C_0) * icf(t_i) * uc(t_i, C_0) * dc(t_i, C_0) \tag{5}$$

この式を 3.1.1 に示した手法で切り出した各語に適用し，地域度の高い語を地域の特徴語として用いる．

3.2 地域ユーザ検索

地域ユーザ検索では 3.1 の手法で抽出された上位の特徴語を多く呟くユーザを地域ユーザと推定する．上位の特徴語から各ユーザ及び対象としている特定地域に対してベクトルを生成した後 (それぞれユーザベクトル，地域ベクトルと呼ぶ．)，ユーザベクトルと地域ベクトルとのコサイン類似度をユーザごとに計算し，類似度の高いユーザをその地域のユーザとする．ただし，計算対象となるユーザは地域特徴語を用いて収集する．類似度の計算式を式 (6) に示す．

$$sim(v_u, c_0) = \frac{\sum_i w_{ui} w_{c_0 i}}{\sqrt{\sum_i w_{ui}^2}\sqrt{\sum_i w_{c_0 i}^2}} \tag{6}$$

ここで $v_u = (w_{u1}, w_{u2}, ...w_{un})$ はユーザ u のユーザベクトル，$c_0 = (w_{c_0 1}, w_{c_0 2}, ...w_{c_0 n})$ は地域 C_0 の地域ベクトルであり，w_{ui} と $w_{c_0 i}$ はそれぞれ各ベクトルにおける上位特徴語 h_i の重みである．式 (6) の値が一定の閾値を超えていればユーザ u を地域 C_0 のユーザと判断する．また，地域ベクトルの各重み $w_{c_0 i}$ には，式 (7) のように対応する上位特徴語 h_i の地域度を割り当てる．

$$w_{c_0 i} = loc(h_i, C_0) \tag{7}$$

評価実験では提案手法を含めた四つの手法で地域ベクトルを作成する．

4. 評価実験

提案手法の有効性を検証するために，地域特徴語の抽出の評価実験と，地域ユーザ検索の予備調査及び評価実験を行った．本実験では対象とする特定の生活地域を京都大学周辺，その周辺の他の地域を京都市の全 11 区（北区，上京区，左京区，中京区，東山区，山科区，下京区，南区，右京区，西京区，伏見区）

表 1　実験データの各地域ごとの Twitter ユーザ数

地域	京大周辺	北区	上京区	左京区	中京区	東山区
ユーザ数	349	389	346	421	399	316

地域	山科区	下京区	南区	右京区	西京区	伏見区
ユーザ数	315	331	335	350	322	365

と設定し，1 地域あたり約 350 から 400 人のユーザの 10 月 1 日から 11 月 30 日までの 2 ヶ月間の Tweet 計 1,843,188 件を実験データとして用いる．尚，京大周辺のユーザには Twitter プロフィールの自己紹介の欄に “京大生”，“京大”，“KU”，あるいは “Silly Fox” などの京都大学固有の団体名，あるいは “総人” などの京都大学特有の学部を示す単語が記載されているユーザ，もしくは Twitter プロフィールの場所の欄に京都大学と記載されているユーザを人手で収集して用いている．各区のユーザについては．Twiiter ユーザをプロフィール情報で検索できるサービスである twpro (注5) と Twitter プロフィール検索 (注6) の API を用いて各区のユーザを取得した．

4.1　地域特徴語抽出の評価

前述した実験データ全てを使って提案手法による特徴語のランキングを行った．手法の評価には，地域度の上位 100 語，500 語，1000 語について nDCG(normalized Discounted Cumulative Gain) を用いる．各語の京都大学周辺との関連度合いは 0 から 3 の 4 段階とし，人手でスコア付けを行った．用いた上位 k 語に対する DCG の式を式 (8) に示す．

$$DCG_k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{\log_2(i+1)} \tag{8}$$

ここで rel_i は第 i 位の語の京都大学周辺との関連度合いを表す．比較手法には，以下の三つの手法を用いる．

- 手法 B1

対象地域での語の出現頻度のみを考慮した式 (9) を用いて語のランキングを行う

$$loc_{B1}(t_i, C_0) = tf(t_i, C_0) \tag{9}$$

- 手法 B2

対象地域での語の出現頻度とその他の地域での出現を考慮した式 (10) を用いて語のランキングを行う

$$loc_{B2}(t_i, C_0) = \log_2(1 + tf(t_i, C_0)) * \log_2(1 + icf(t_i)) \tag{10}$$

- 手法 B3

Paik の研究 [12] で紹介されている RITF(Relative Intra-document TF) を，提案手法の地域度 (式 (5)) に $rtf(t_i, C_0)$ の代わりに使う式 (11) を用いて語のランキングを行う

$$loc_{B3}(t_i, C_0) = ritf(t_i, C_0) * icf(t_i) * uc(t_i, C_0) * dc(t_i, C_0) \tag{11}$$

(注5)：http://twpro.jp/

(注6)：http://tps.lefthandle.net/

表 2　手法ごとの地域度による語のランキング TOP20．太字は京都大学周辺に関連する語を表す

順位	提案手法	手法 B1	手法 B2	手法 B3
1	**総人**	今日	ミニスイーツ局	今日
2	**NF**	時間	ミニスイーツ	時間
3	**メディセン**	明日	ライトブースト	明日
4	ナイスロック	みたい	オポ	みたい
5	ナイスナイス	自分	スイーツ局	自分
6	**京大**	京都	ライトブー	京都
7	**クラシス**	好き	京大マップ	最近
8	**授業**	ちゃん	カーリング	好き
9	バイト	**大学**	ナイスロック	**京大**
10	**履修**	感じ	回演奏会	みんな
11	**共北**	**授業**	ナイスナイス	感じ
12	**実験**	**京大**	writeboost	ところ
13	**D 進**	バイト	**KU 界隈**	ちゃん
14	**ルネ**	最近	山崎はるかさん	**大学**
15	ロックスター・エナジードリンク	ところ	**KU 界隈 B**	**授業**
16	**4 共**	www	界隈 B	バイト
17	**時計台**	みんな	z カラオケオフ	**NF**
18	**回生**	日本	ひなすき	昨日
19	**レポート**	問題	梅子氏	意味
20	英語	意味	カドラ	日本

表 3　手法ごとの語のランキングに対する NDCG の評価値

語数＼手法	提案手法	手法 B1	手法 B2	手法 B3
上位 100 語	**0.884**	0.429	0.371	0.415
上位 500 語	**0.836**	0.462	0.468	0.469
上位 1000 語	**0.801**	0.472	0.523	0.490

ここで，RITF は対象地域語での語の出現頻度をその地域の語の出現頻度の平均で割って正規化したものである，RITF の計算式を式 (12) に示す．ただし $|T_{C_0}|$ は地域 C_0 のユーザが呟いた Tweet から切り出した語の語彙数とする．

$$ritf(t_i, C_0) = \frac{\log_2(1 + tf(t_i, C_0))}{\log_2(1 + \frac{1}{|T_{C_0}|} \sum_j tf(t_j, C_0))} \tag{12}$$

結果として表 2 に各手法ごとのランキングの一部（上位 20 件）を，表 3 に各手法ごとの上位 100 語，500 語，1000 語に対する nDCG による評価値を示した．表 2 から提案手法によるランキングでは京都大学周辺に関連する言葉（総人，NF，メディセン，クラシス，共北.. 等）が多く上位にランクされることが分かった．これに対し，手法 B1，B3 によるものは一般的な語 (今日，時間，明日，みたい，好き，感じ.. 等) が上位に多く見られ，手法 B2 によるものには一般的ではないが京都大学周辺とは関連のない語（ミニスイーツ，ミニスイーツ局，ライトブースト，オポ.. 等）が上位にランクされることが分かった．また表 3 の通り，提案手法によるランキングは上位 1000 語に対する nDCG の評価値が 0.801 と他の手法に比べておよそ 0.3 も上回っている．これらの結果から，提案手法が地域特徴語の抽出に有効であると考えられる．

以下で各手法についての考察を述べる．

表 4　手法 P1 のしきい値と再現率，適合率，F 値の関係

しきい値	再現率	適合率	F 値
0.65	0.281	0.671	0.395
0.60	0.341	0.575	0.428
0.55	0.418	0.495	0.453
0.50	0.501	0.422	**0.459**
0.45	0.573	0.380	0.457
0.40	0.639	0.329	0.434
0.35	0.716	0.298	0.421

表 5　手法 P2 のしきい値と再現率，適合率，F 値の関係

しきい値	再現率	適合率	F 値
0.40	0.344	0.732	0.468
0.375	0.438	0.709	0.540
0.35	0.504	0.624	0.558
0.325	0.567	0.559	**0.563**
0.30	0.650	0.477	0.550
0.275	0.716	0.419	0.529
0.25	0.751	0.365	0.492

表 6　手法 P3 のしきい値と再現率，適合率，F 値の関係

しきい値	再現率	適合率	F 値
0.25	0.183	0.831	0.300
0.225	0.309	0.755	0.439
0.2	0.410	0.598	0.486
0.175	0.553	0.512	**0.532**
0.15	0.659	0.399	0.497
0.125	0.734	0.300	0.426
0.1	0.814	0.227	0.355

- 手法 B1

手法 B1 は単純に語の出現頻度を指数としているため一般的に多く呟かれる語が上位にランクされたと考えられる．手法 B1 の上位には “京大” や “大学” などの京都大学周辺に関連のある語もランクされているが，これらは京大生の Tweet において出現頻度を高いから上位にランクされているいるだけで，出現頻度の低い京都大学に関連のある語は下位にランクされてしまった．

- 手法 B2

手法 B2 は，手法 B1 とは違って式 (2) の効果で他の地域で呟かれない語を上位にランクしているが，“ミニスイーツ” や “ライトブースト” と京大に関連のない語を上位にランクしてしまっている．詳しく調べると，これらの上位の語の多くは 1 人あるいは少数のユーザだけが何度も呟いている語や短期間に多く呟かれた語であった．このため，他の地域には出現しないが地域に関連のない語が上位にランクされることがあるとわかる．

- 提案手法と他の手法の比較

手法 B1，B2 に対し，提案手法では式 (3) でユーザ数を，式 (4) で語が呟かれる日数を考慮に入れているため，より優れた結果が得られたと考えられる．また，手法 B3 でも式 (3),(4) を組み込んでいるが，提案手法の式 (1) が他の地域の出現頻度を考慮して語の相対的な出現頻度を出すのに対して，RITF は京大周辺の Tweet 内での出現頻度だけを考慮して出現頻度の正規化を行っていることから，手法 B1 と似た結果になったと考えられる．このことから式 (1) も特徴語抽出に有効であるとわかった．

しかしながら，今回の評価実験では京都大学周辺のユーザとして京大生を選んでいるので，このように提案手法が優れた結果を得られたのは “京大生” という共通の属性を持っていたからだと考えられる．また今回の nDCG の評価では関連度合いを 1 人で付けたため，今後より多くのユーザに．評価してもらう必要があると考えられる．

4.2　地域ユーザ検索の評価

提案手法の有効性を確かめるために地域特徴語を利用した地域ユーザ検索の評価実験を行った．本節では，まず地域ユーザの分類のための適切なしきい値を決定する予備調査について述べ，次に地域ユーザ検索の評価実験について述べる．

4.2.1　予 備 調 査

地域ユーザ検索手法の予備調査として，全ユーザの Tweet から提案手法を用いて京都大学周辺の上位 1000 語の特徴語を特徴量とする地域ベクトルを作成し，この地域ベクトルと全地域の各ユーザの Tweet 群から作成したユーザベクトルのコサイン類似度を計算してユーザの分類を行った．ユーザ u のユーザベクトル $v_u = (w_{u1}, w_{u2}, ...w_{u1000})$ の特徴量も京都大学周辺の上位 1000 語の特徴語とし，重みは以下の三通りの手法で作成する．

- 手法 UV_b

ユーザが語を含む Tweet を一回以上すれば重みを 1 に，そうでなければ 0 にする

$$w_{ui} = \begin{cases} 0 & \text{if } utf(h_i, U) = 0 \\ 1 & \text{else} \end{cases}$$

- 手法 UV_f

ユーザが語を含む Tweet を呟いた回数を重みとする

$$w_{ui} = utf(h_i, u)$$

- UV_{fd}

ユーザが語を含む Tweet を呟いた回数とユーザが語を含む Tweet を呟いた日数の積を重みとする

$$w_{ui} = utf(h_i, u) * \frac{udc(h_i, u)}{|D|}$$

ここで h_i は重み w_{u1} に対応する語を表し，$utf(h_i, u)$ はユーザ u が語 h_i を含む Tweet を呟いた回数，$udc(h_i, u)$ はユーザ U が語 h_i を含む Tweet を呟いた日数を示す．予備調査では以下の三通りの手法でユーザ分類を行った．

- 手法 P1

手法 UV_b により重みを作成した各ユーザベクトルと地域ベクトルとの類似度による分類

- 手法 P2

手法 UV_f により重みを作成した各ユーザベクトルと地域ベクトルとの類似度による分類

- 手法 P3

手法 UV_{fd} により重みを作成した各ユーザベクトルと地域ベクトルとの類似度による分類

それぞれの手法での，しきい値と再現率 ($Recall$)，適合率 ($Precision$)，F 値の関係を表 4,5,6 に示す．ただし F 値は以下の式 (13) で計算を行った．

$$(\text{F 値}) = \frac{2Recall * Precision}{Recall + Precision} \tag{13}$$

F 値の最高値から手法 P2 が最もユーザ分類に適していると予想される．手法 P1 では (しきい値)=0.50，手法 P2 では (しきい値)=0.325，手法 P3 では (しきい値)=0.175 で F 値が最も大きくなることが分かったので，これらのしきい値を用いて次節で地域ユーザ検索の評価実験を行う．

予備調査では本実験で用意した全ユーザの Tweet を使って地域ベクトルを作成し，全ユーザのユーザベクトルとの類似度を計算してユーザ分類を行った．これに対し評価実験では全ユーザを評価ユーザと訓練ユーザの 2 グループに分け，訓練ユーザの Tweet を使って地域ベクトルを作成し，評価ユーザのユーザベクトルとの類似度を計算してユーザ分類を行う．

4.2.2 評価実験

地域ユーザ検索手法の評価をするために，本実験で用意した全ユーザのうち各地域 100 人計 1200 人を評価ユーザ，それ以外のユーザを訓練ユーザとして評価実験を行った．3 章で述べたように，本研究では，基本的にユーザベクトルと地域ベクトルのコサイン類似度に基づいて地域ユーザの識別を行う．ユーザベクトルと地域ベクトルの構築手法は以下に示す通り，それぞれ三通りと四通りがある．これらの組み合わせで得られた十二通りのユーザ分類手法（表 7，表 8 を参照）の比較を行った．

ユーザベクトルの構築手法は，4.2.1 節で述べた,UV_b，UV_f，UV_{fd} の三通りである．地域ベクトルの構築手法は以下に挙げる四通りである．

- 手法 CV_{loc}

表 7　評価実験における各手法の概要

手法	ユーザベクトルの特徴量	地域ベクトルの特徴量	地域ベクトルの重み
CV_{loc}	地域特徴語上位 1000 語		対応する地域度
CV_{lua}			
CV_{B1}	訓練ユーザ中の京都大学周辺のユーザの Tweet で出現頻度の高い上位 1 万語		訓練ユーザ中の京都大学周辺のユーザのユーザベクトルの平均
CV_{B2}	各ユーザの Tweet ごとに出現頻度の高い上位 1000 語	訓練ユーザ中の京都大学周辺のユーザのユーザベクトルの特徴量全て	

表 8　各ユーザ分類の再現率，適合率と F 値

手法	再現率	適合率	F 値
$CV_{loc}+UV_b$	0.560	0.403	**0.469**
$CV_{lua}+UV_b$	0.550	0.320	0.404
$CV_{B1}+UV_b$	0.550	0.314	0.400
$CV_{B2}+UV_b$	0.540	0.195	0.286
$CV_{loc}+UV_f$	0.580	0.542	**0.560**
$CV_{lua}+UV_f$	0.510	0.395	0.445
$CV_{B1}+UV_f$	0.510	0.481	0.495
$CV_{B2}+UV_f$	0.570	0.291	0.385
$CV_{loc}+UV_{fd}$	0.450	0.441	**0.446**
$CV_{lua}+UV_{fd}$	0.630	0.251	0.359
$CV_{B1}+UV_{fd}$	0.50	0.379	0.431
$CV_{B2}+UV_{fd}$	0.690	0.220	0.334

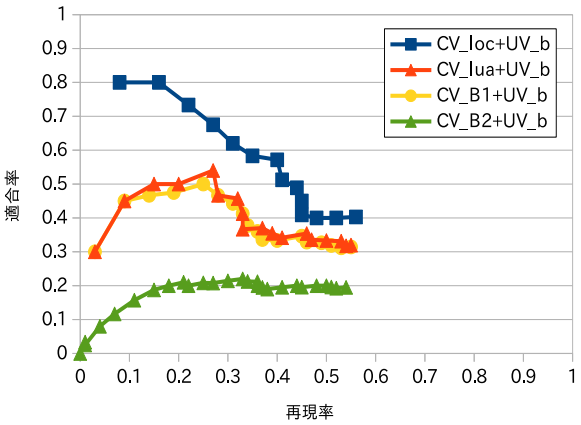

図 1　各地域ベクトル作成手法と UV_b を用いた際の再現率と適合率の PR 曲線

訓練ユーザの Tweet から抽出された京都大学周辺の特徴語の上位 1000 語を特徴量とする地域ベクトルを生成する．それぞれの語の重みは，特徴語を抽出する際に計算した類似度とする．

- 手法 CV_{lua}

訓練ユーザの Tweet から抽出された京都大学周辺の特徴語の上位 1000 語を特徴量とする地域ベクトルを生成する．それぞれの語の重みは，訓練ユーザ中の京都大学周辺のユーザのユーザベクトルにおけるその語の重みの平均とする．つまり以下の式 (14) で地域ベクトルのそれぞれの語の重みを計算する．ただし $c_0 = (w_{c_0 1}, w_{c_0 2}, ...w_{c_0 n})$ は地域ベクトル，$v_l = (w_{l1}, w_{l2}, ...w_{ln})$ は訓練ユーザ中の京都大学周辺のユーザのユーザベクトル，$|L_{C_0}|$ は訓練ユーザ中の京都大学周辺のユーザの総数を表す．

$$w_{c_0 i} = \frac{1}{|L_{C_0}|} \sum_{j=1}^{|L_{C_0}|} w_{l_j i} \tag{14}$$

- 手法 CV_{B1}

訓練ユーザ中の京都大学周辺のユーザの Tweet で出現頻度の高い上位 10000 語を特徴量とする地域ベクトルを生成する．それぞれの語の重みは訓練データ中の京都大学周辺のユーザのユーザベクトルにおけるその語の重みの平均とする．

- 手法 CV_{B2}

訓練ユーザ中の京都大学周辺の各ユーザの Tweet ごとに出現頻度の高い上位語 1000 を集めて，地域ベクトルを作成する．それぞれの語の重みは訓練データ中の京都大学周辺のユーザのユーザベクトルにおけるその語の重みの平均とする．

ただし，CV_{loc}，CV_{lua}，CV_{B1} では地域ベクトルとユーザベクトルの特徴量は同じだが，CV_{B2} でのユーザベクトルの特徴量は各ユーザの Tweet 群ごとに出現頻度の高い上位 1000 語である．

各分類手法について，しきい値は，地域ベクトル作成法に手法 CV_{loc} を用いるユーザ分類手法については予備調査で求めた値を，それ以外のユーザ分類手法については F 値が最も高くなる値を採用した．

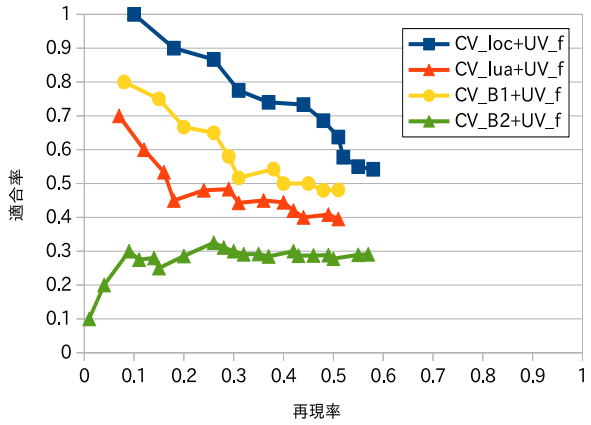

図 2　各地域ベクトル作成手法と UV_f を用いた際の再現率と適合率の PR 曲線

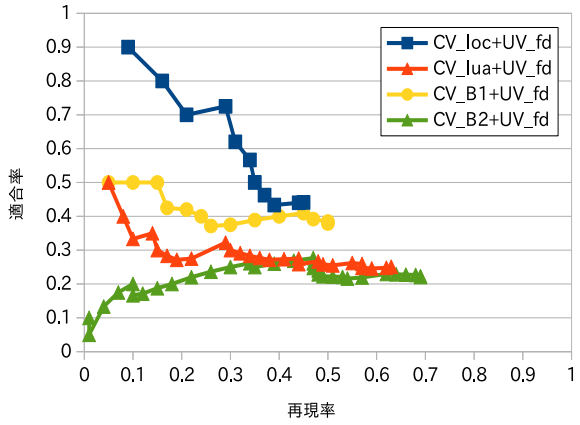

図 3　各地域ベクトル作成手法と UV_{fd} を用いた際の再現率と適合率の PR 曲線

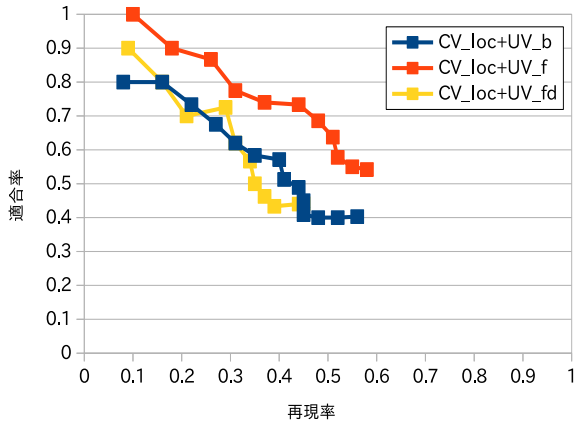

図 4　提案手法と各ユーザベクトル重み作成法を用いた際の再現率と適合率の PR 曲線

結果として，表 8 に各手法によるユーザ分類の再現率，適合率，F 値を，図 1，2，3 にユーザベクトルの重み作成法にそれぞれ UV_b,UV_f,UV_{fd} を用いた場合の各手法によるユーザ分類の再現率と適合率の PR 曲線を，図 4 に提案手法に UV_b,UV_f,UV_{fd} を用いたユーザ分類の再現率と適合率の PR 曲線を，図 5 に全ユーザ分類の再現率と適合率の PR 曲線を示す．ただしこれらの図 1~5 は，しきい値によって分類したユーザのうち，類似度の高い順にユーザを 10 人ずつ増やしていって再現率，適合率を計算しプロットしたものである．

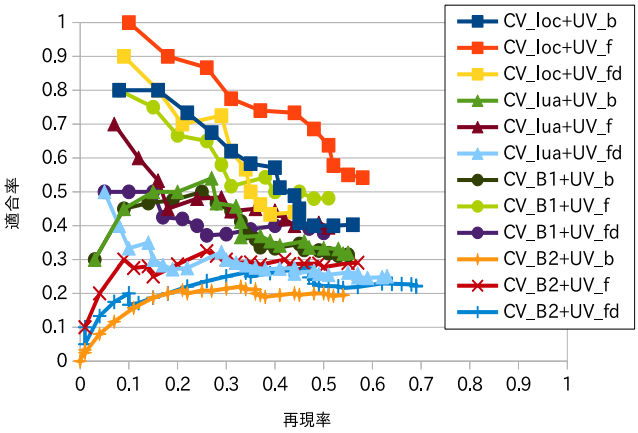

図 5　全ユーザ分類の再現率と適合率の PR 曲線

ユーザベクトル作成法が同じユーザ分類手法を比較すると，図 1，2，3 より，いずれのユーザベクトル作成法を使う場合でも，ユーザ分類の性能は地域ベクトル作成法に手法 CV_{loc} を使ったものが最も高かったことが分かる．

次に地域ベクトル作成法が CV_{loc} の各分類手法をユーザベクトル作成方法で比較すると，図 5 から，手法 UV_f を用いる場合にユーザ分類の性能が最も高いことが分かる．全手法で比較すると，ユーザ分類手法に CV_{loc}+UV_f を用いる場合が，表 7 より F 値が 0.560 と最も高く，図 6 から見てもユーザ分類の性能が最も高いことが分かった．これらの結果について以下で考察を述べる．

- 手法 CV_{loc} と CV_{lua} の比較

地域ベクトルおよびユーザベクトルの特徴量が上位特徴語 1000 語である，提案手法 CV_{loc} と手法 CV_{lua} を比較すると，ユーザベクトルの重み作成法が UV_b，UV_f，UV_{fd} のいずれの場合でも CV_{loc} の方が性能が高かったことから，地域ベクトルの重みに訓練ユーザ中の京都大学周辺のユーザのユーザベクトルの重みの平均を用いるよりも，提案した地域度を用いた方が良いと考えられる．

- ユーザベクトル作成法の比較

どのユーザ分類手法においても，ユーザベクトル作成法ごとの性能を比較すると UV_f を用いた場合が最も F 値も高く性能が良いことが分かった．これに関して，UV_b は一度でも語を呟けばその語の重みが 1 になるので，ユーザ間の重みの差が狭まってしまって分類の性能が UV_f の場合よりも悪くなったと考えられる．

次に UV_{fd} について，地域特徴語の指数に組み込まれている式 (4) が対象地域のユーザの少なくとも 1 人が語を呟いた日数の割合なのに対し，UV_{fd} ではそのユーザ個人が呟いた日数の割合を考慮している．このため，対象地域で呟かれた日数の多

い地域特徴語でも，ユーザが一回しか呟いていなければユーザのユーザベクトルにおけるその地域特徴語の重みは小さくなってしまう．一方 “今日” などの地域で出現頻度が高く出現日数も多い一般的な語は，ユーザ個人も呟く回数と日数が多く，ユーザベクトルの語に対する重みが大きくなる．結果として類似度の計算が一般的な語に強く影響されることになり，対象地域のユーザと他の地域のユーザの分類が難しくなったのだと考えられる．

5. おわりに

本論文では Twitter から日常生活に役立つ生活地域の情報を発見するための情報源として，その生活地域に関わりのあるユーザの Tweet が有効な情報源であると考え，そのような生活地域の Twitter ユーザを，語の出現頻度，語を発信するユーザの数，語が発信された日数，生活地域の周辺のその他の地域との差異を考慮して抽出した特徴語と，その特徴語を抽出した指標である地域度を用いて検索する手法を提案した．評価実験の結果，地域特徴語の抽出の提案手法は，上位 1000 語に対する nDCG の評価値が 0.801 と比較手法を大きく上回る有効性を示した．地域ユーザ検索については，ユーザ分類の評価実験の結果，地域特徴語の上位語を地域ベクトルとユーザベクトルの特徴量とし，地域度を地域ベクトルの重みに用いて地域ベクトルとユーザベクトルのコサイン類似度を計算する提案手法が比較手法を上回る性能を示し，特にユーザベクトルの重みを各ユーザの Tweet における語の出現頻度とした場合は，(F 値)=0.560 と今回行ったユーザ分類の中で最も高い性能を示した．これらの結果から，地域度による地域特徴語の抽出およびそれに基づく地域ユーザの検索が有効であることが分かった．今後、以下の課題について検討する予定である．

- 地域特徴語抽出の評価

本研究の地域特徴語の nDCG の評価は，関連度合いを 1 人で付けたために提案手法の評価値が高くなった可能性がある．これについて完全な正解データを用意することはできないので，多数の京大生にアンケートを取って関連度合いの正解データを作り，できる限り正しい評価値を出す予定である．

- 実験データの偏り

本研究では，対象とする生活地域の京都大学周辺のユーザとして京都大学の学生を用いている．このため，対象地域のユーザが “京大生” という共通の属性を持っていたために，地域特徴語の抽出と地域ユーザの分類が提案手法が有効性を示した可能性がある．このことを検証するために，今後別の地域についても対象とする生活地域として提案手法を用いて評価を行う予定である．また，“学生” という属性に着目して，異なる大学に対して本研究と同様の評価実験を行うことや，異なる大学間同士で評価実験を行うことも検討している．

- Twitter からの地域ユーザの検索

地域ユーザ検索の評価実験では，今回はあらかじめ評価データを用意してユーザ分類を行ったが，実際にユーザ検索をする場合には Twitter から評価するユーザを何らかの手段で選ぶ必要がある．同じく訓練データとなるシードユーザについても今回はあらかじめ用意したが．対象地域の範囲が狭ければ狭いほどシードユーザとなるユーザも収集が困難になる，これら問題に付いても今後の課題として研究を行っていく．

- 有益な日常情報の発見

本研究で提案した小さい生活地域のユーザ検索手法は，この生活地域のユーザから日常情報に役立つ情報を発見するためのものである．今後，この小さい範囲のユーザから有益な日常情報を発見する方法について研究を行っていく．

謝　辞

本研究の一部は，科研費（課題番号 25700033) と SCAT 研究費助成による．

文　献

[1] T.Sakaki, M.Okazaki, and Y.Matsuo,: Earthquake Shakes Twitter Users:Real-time Event Detection by Social Sensors, *Proceedings of the 19th International Conference on World Wide Web*, pp. 851–860 (2010).

[2] 榊剛史, 松尾豊: ソーシャルメディアからの人物目撃情報抽出システムの試作, 第２５回人工知能学会全国大会 (JSAI2011) 予稿集 (2011).

[3] 土屋圭, 豊田正史, 喜連川優: マイクロブログを用いた鉄道の運行トラブル状況抽出に関する一検討, 信学技報, Vol. 113, No. 150, pp. 175–180 (2013).

[4] 松村飛志, 安村通晃: 街に着目した Twitter メッセージの自動収集と分析システムの提案と試作, 情報処理学会インタラクション 2010 予稿集 (2010).

[5] 渡辺一史, 大知正直, 岡部誠, 尾内理紀夫: Twitter を用いた実世界ローカルイベント検出, 第 4 回楽天研究開発シンポジウム予稿集 (2011).

[6] 杉谷卓哉, 白川真澄, 原隆浩, 西尾章治郎: 位置情報付きツイートの時空間的局所性の解析によるローカルイベント検出手法, DICOMO2012 シンポジウム 予稿集, pp. 1704–1711 (2012).

[7] Z.Cheng, J.Caverlee and K.Lee, : You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users, *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp. 759–768 (2010).

[8] 堂前友貴, 関洋平: 地域に偏りのあるトピックを用いた Twitter ユーザの生活に関わる地域推定, 情報処理学会・情報学基礎研究会報告, Vol. 2013, No. 8, pp. 1–6 (2013).

[9] 西村駿人, 数原良彦, 鷲崎誠司: 地域特徴語選択を用いたマルチクラス分類による Twitter ユーザの居住地推定, 信学技報, Vol. 112, No. 367, pp. 23–27 (2012).

[10] 長谷川馨亮, 馬強: Twitter からの地域特徴語辞書の構築とその観光情報への応用, 第６回データ工学と情報マネジメントに関するフォーラム（DEIM2014）予稿集, (2014).

[11] C.Li, A.Sun, J.Weng and Q.He: Exploiting Hybrid Contexts for Tweet Segmentation, *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 523–532 (2013).

[12] J.H.Paik: A Novel TF-IDF Weighting Scheme for Effective Ranking, *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 343–352 (2013).