位置情報付き SNS データによる空間スケールに着目した地域特徴語抽出

秋庭　武・藤田　秀之・大森　匡・新谷　隆彦

Extracting Local Feature Words Focusing on Spatial Scale Change

Takeru AKIBA, Hideyuki FUJITA, Tadashi OHMORI, Takahiko SHINTANI

Abstract: In this paper, we propose a method to extract characteristic words (local feature word) for each region from SNS data with local information, considering scale. For example, the local feature words of the target area in Ueno Park are “Taito-ku” and “Ueno” on the Tokyo scale (characterizing the target area in Tokyo), but “Gorilla” and “Panda” on the Ueno Park scale (characterizing the target area in Ueno Park). Such a framework for extracting different local feature words is considered useful. Therefore, we propose a method to extract words that characterize the target area at a specified scale as regional feature words. We analyze the proposed method and evaluate its effectiveness, and examine the possibility of a new regional classification method using the proposed method.

Keywords:地域特徴語（local feature words），可視化（visualization），スケール（scale）

1. 背景と目的

地理的な位置や場所に関する情報を持つデータ(空間データ)が大量に流通しており，地域や場所ごとに,特徴や話題,イベント等を抽出し,地図上に可視化するシステムは盛んに研究されている．中でも，Twitter を代表とする SNS の位置情報付き投稿データを用い，地域ごとに特徴的な語(地域特徴語)を抽出し，地図上に可視化する試みは盛んである(例えば[1])．地図作成において,提示する情報の詳細さをスケール(対象領域の広さ)に応じて適切に定めることは，基本的要件であり，人手による調整は欠かせない．地図上のテキストラベルも，詳域のスケールの地図では，広域の地図と比較し，より詳細な内容となる．例えば，上野公園のガイドマップに「東京都」「台東区」等のテキストラベルは用いられない．しかし，SNS からの地域特徴語抽出において，こうした点は考慮されてこなかった．そこで本研究では，スケールの変化に応じた語の一般性の変化に着目した地域特徴語の新しい抽出・可視化手法に向けて，語の重要度として tfidf 値を用いる際に，語の一般性を算出するスケール(領域の広さ)を変化させる枠組みの有効性について分析を行う．

2. SNS からの地域特徴語抽出手法

ツイート集合からの地域特徴語抽出において，一般的に用いられる手法(初等的手法と呼ぶ)を説明する．対象データは，位置情報付きツイートpの集合である．$p = \langle loc, text \rangle$と定める．$p.loc = (x, y)$は地理空間上の位置座標$(x, y)$として与えられる位置情報である．$p.text$は，最長 140 文字のテキストである．各ツイートのテキストを形態素解析し，語の頻度付き集合として扱う．地理空間をグリッドで分割し，グリッドの空間セルごとに,位置情報が空間セル内に含まれる全てのツイートを集約し，ひとつの文書とみなす．文書は，その文書における語の重要度を要素とするベクトルとして扱われる．ベクトルの次元数は，全文書中に出現する語の種類数であり，各次元は各語に対応する．文書の特徴語とは，各文書において重要度が高い語である．重要度として，文書中での出現頻度や tfidf 値が用いられる．一般に，重要度について，指定された上位k件のリストを，特徴語リストとして抽出する．文書が地域に対応する場合，特徴語は地域特徴語と呼ばれる．

秋庭　武，藤田　秀之，大森　匡，新谷　隆彦

電気通信大学院情報理工学研究科

E-mail: akiba@hol.is.uec.ac.jp

3. 提案手法

3.1. ねらい

tfidf 値は，文書の集合を入力とし，各文書における各語に対して算出される．文書中での出現頻度が高い語の重要度を上げるが，他の多くの文書に出現する語，すなわち，一般性が高い語の重要度を相対的に下げる．語の一般性は，全文書のうち，その語が出現する文書数を用いて算出される．すなわち，語の一般性は全文書を対象に算出される．他方で，文書があらかじめカテゴリに分類されている場合，全文書ではなく，カテゴリ内の文書集合を対象に語の一般性を算出する手法も用いられる．同手法による特徴語は，カテゴリ内の各文書の弁別性が高いことが期待できる．理由として，全文書での一般性は高くないが，指定されたカテゴリ内での一般性は高い語の重要度が相対的に下がるためである．語の一般性を，全文書ではなく，カテゴリ内の文書を用いて算出することは，地域特徴語の抽出において，語の一般

性を，特定の地理的領域内に含まれる文書集合を用いることに対応する．本研究では，語の一般性の算出に用いる領域を，コンテクスト領域と呼ぶ．コンテクスト領域を多段階的に変化させることで，コンテクスト領域内で一般性が高い語をフィルタリングし，抽出される地域特徴語と重要度の遷移を分析する．

先行研究として，Feick らの研究[2]では，全国スケールの tfidf 値を都市スケールの tfidf 値で除した値を LG 比として，地域の固有性を示す一指標としている．また，鈴木の研究[3]では，LG 比が観光スポットの固有性を示す指標になりうる可能性を示している．

3.2. 語の一般性の算出に用いる領域を考慮した重要度

グリッドの単位領域である空間セルcで構成される矩形領域として，いくつかの地理的な領域(図 1)を定義する．全データ領域Dは，対象データのすべての位置座標を含む最小の矩形領域であり，固定の領域である．コンテクスト領域$C(C \subseteq D)$は，後述する地域特徴語の抽出において，語の一般性(idf 値)を算出する対象とする領域である．

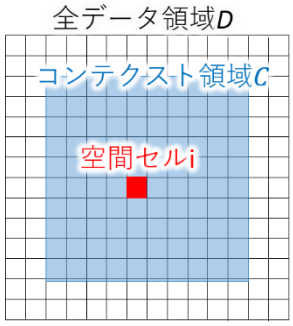

図 1 本研究で用いる領域の定義

コンテクスト領域がCのとき，空間セルiにおける，語wの重要度$d_{w,i,C}$を，以下のように定義する．

$$\boldsymbol{d_{w,i,C} = tf_{w,i} \cdot idf_{w,C} = \frac{fr_{w,i}}{\sum_{v \in i} fr_{v,i}} \cdot \log\left(\frac{|C|}{df_{w,C}}\right)} \quad (1)$$

ここで，$tf_{w,i}$は空間セルiにおける単語wの出現頻度$fr_{w,i}$を，iに出現する全ての語の出現頻度の和で除した値である．$fr_{v,i}$は空間セルiにおける語$v \in i$の出現頻度である．$idf_{w,C}$は，領域C内で一般的な語の重要度を下げるフィルタである．$df_{w,C}$は，Cに含まれ，語wが出現する空間セルの数である．初等的手法における tfidf 値は，$d_{w,i,C}$においてCを全データ領域Dとしたものである．

続いて，以降に示す基準語を用いた重要度の正規化を行う．基準語$base_i$を，対象とする空間セルiにのみ出現し，他のセルには存在しない人工的な語と定義する．$base_i$は，任意のCに対して，i内で局所性がもっとも高い(もっとも一般性が低い)語となる．式(1)による$base_i$の重要度は以下となる．

$$\boldsymbol{d_{base_i,i,C} = tf_{base_i,i} \cdot idf_{base_i,C} = \frac{fr_{base_i,i}}{\sum_{v \in i} fr_{v,i}} \cdot \log(|C|)} \quad (2)$$

上記を用い，基準語$base_i$により正規化した語wの重要度$d'_{w,i,C}$を，以下のように定義する．

$$\boldsymbol{d'_{w,i,C} = \frac{d_{w,i,C}}{d_{base_i,i,C}} = \frac{d_{w,i,C}}{tf_{base_i,i} \cdot \log(|C|)}} \quad (3)$$

ここで，空間セルiを固定して，コンテクスト領域Cと語wを変数とする場合，$tf_{base_i,i}$は，各重要度$d'_{w,i,C}$に対して定数の係数となる．

4. 実験

4.1. データセットと方法

コンテクスト領域の変化に応じた語の重要度の遷移の分析を目的とした実験を行う．グリッドの空間セルを約 1km 四方の矩形とする(総務省標準地域・3 次メッシュ)．データとして，位置情報付きツイート約 340 万件(2018 年 7 月 1 日から 2 ヶ月間分)を用いる．複合語に対応した名詞のみを対象とする．本文に含まれる外部アプリ等による自動入力部分は削除する．同じユーザーの多頻度投稿による好ましくない影響を除くために，同じユーザーからは，同じ空間セルでは同じ語を複数回カウントしないように集計する．コンテクスト領域の大きさ(スケール)を，対象空間セル周辺の領域から全データ領域へと段階的に拡大し，各スケールにおける上位 20 件の地域特徴語と重要度を算出する．

4.2. 結果と重要度の遷移による分類

ひとつの空間セルに関して，横軸をコンテクスト領域内の空間セル数(スケール)，縦軸を各スケールにおける各語の正規化した重要度とし，その遷移を上野公園周辺の空間セルを例として図 2 に示す．図中には各スケールで一度でも上位 20 件に出現した全ての語の遷移を表示している．図中の赤点は，遷移の標準偏差を閾値として抽出した重要度の極大値であり，後述するピークが存在することを示す．tf 値はスケールによらず一定なため，重要度の遷移は，基準語の idf 値に対する各語の idf 値の比率の遷移に従う．したがって重要度が増加している範囲では，基準語に対する比として，各語の局所性が増加している．この場合，増加した空間セル群中，各語を含む空間セルが少なかったことを別の実験で確認している．スケールと重要度の遷移に基づいた地域特徴語群の抽出に向け，表 1 に示すとおり，図 1 のグラフ右端で上位 20 件の語を重要度の遷移に基づき，ほぼ増加している遷移(P 群)，ほぼ一定である遷移(Q 群)，ほぼ減少あるいは下に凸である遷移(R 群)，ピークがある遷移(S 群)に分類した．

表 2 語の重要度の遷移による地域特徴語の分類

群	重要度の遷移	地域特徴語
P 群	ほぼ増加	上野，エッシャー展，ミラクルエッシャー展，上野公園，アメ横，不忍池，花，今日，御徒町
Q 群	ほぼ一定	アメヤ横丁，シャンシャン
R 群	ほぼ減少(下に凸)	夜
S 群	ピークあり	上野動物園，蓮，エッシャー，森美術館，パンダ，動物園，上野恩賜公園，分待ち

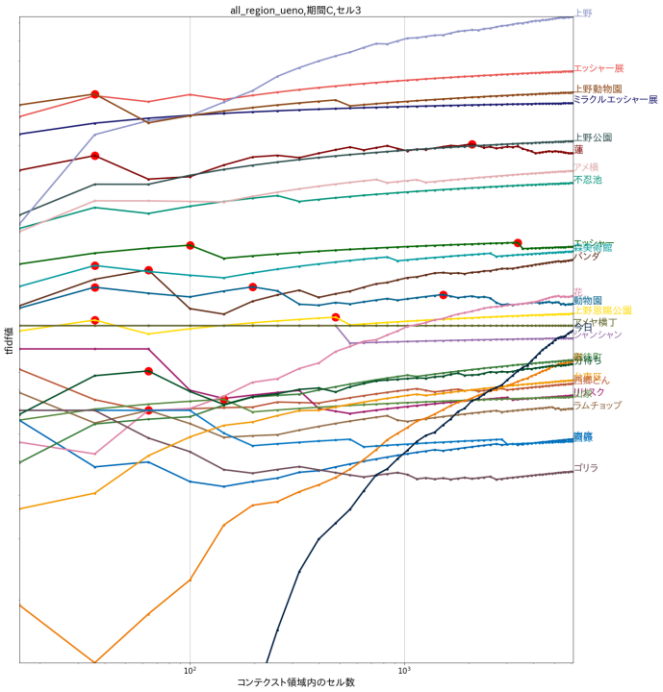

図 2 地域特徴語群の正規化した重要度の遷移

5. 評価と考察

重要度の遷移による語の分類の妥当性を，他の自明な指標による分類にどの程度対応付けられるか，という点に基づき評価する．自明な指標として，名詞の種類(普通名詞か固有名詞)を用い，表 1 の各語を分類し，表 2 に示す．表 1 の各分類項目中の語群のうち，表 2 の各分類項目中の語群に含まれる語の割合を表 3 に示す．数値は実験対象領域から無作為に選んだ 4 セルの平均である．例えば，「P→固」は P 群の語に含まれる固有名詞の割合を示す．本結果より，少なくとも 58%以上の語で，重要度の遷移による語の分類が，自明な分類に対応付けられた．

表 2 名詞の種類による地域特徴語の分類

名詞の種類	地域特徴語
固有名詞	エッシャー展，ミラクルエッシャー展，上野公園，アメ横，不忍池，アメヤ横丁，シャンシャン，上野動物園，エッシャー，森美術館，上野恩賜公園，上野，御徒町
普通名詞	花，今日，蓮，パンダ，動物園，夜，分待ち

表 3 語の重要度の遷移による地域特徴語の分類と別の指標の関係

P→固	Q→普	R→普	S→固
0.8819	0.725	0.5833	1

続いて各群の遷移について考察する．P 群は固有名詞と高い関係がある．固有名詞は一つの意味(場所)を表す語である故に，全国的に対象セル周辺でしか使用されない語が多く含まれると考えられる．Q 群も同様の性質を持つが，ピークがないことは基準語の idf 値と同じ局所性を持つことを示すので,その性質はさらに強いといえる.R 群は普通名詞と高い関係がある．普通名詞は様々な意味で使用される語である故に，対象セル以外の地域でも使用される語が多く含まれると考えられる．S 群も普通名詞と高い関係があるが，重要度のピークは，ピークまでは局所性が高く，以降は一般性が高くなることを示す．従って，普通名詞ではあるが，ピークまでのスケールでは固有名詞的に語が使用されると考えられる.例えば,S 群の「蓮」は，ピークまでのスケールで「不忍池の蓮」という固有名詞的に使用されていると考えられる．ピークまでのスケールでは他の地域で「蓮」があまり出現しないからである．

6. 提案手法を用いた地域分類の検討

提案手法を用いた地域分類の検討を行った．図 3 は上野公園を含む総務省標準地域・2 次メッシュ 1 セルの中の各空間セルに対して，4 章と同様の手法を用いて地域特徴語群の重要度の遷移を求めた結果である．観察すると以下のことがわかる．

1. 地域特徴語群のうち P 群 Q 群の割合が大きい地域がある．(地域 A)
2. 地域特徴語群のうち R 群 S 群の割合が大きい地域がある．(地域 B)
3. 地域特徴語群の重要度の遷移のばらつきが大きい地域がある．
4. 地域特徴語群のうち P 群だけ特に重要度が大きい地域がある．
5. スケールの拡大途中から多くの地域特徴語がグラフ中に現れる地域がある．

図 4 に地域 A,B の代表例を示す．図 4 上段は皇居東御苑周辺の地域特徴語群の重要度の遷移であり，「皇居」「長和殿」「大手門」等の語が抽出されており，地域 A の代表例である．地域 A は，P 群，Q 群の語の割合が大きい地域である．名詞の種類との関連性を考慮すると，抽出される語に固有名詞が多いことを示し，観光スポットのような，特に人気であると予想される地域が多く含まれる．図 2 の上野公園周辺も地域 A に分類される．図 4 下段は葛飾区立石周辺の地域特徴語群の重要度の遷移であり，「立石」「梅割り」「チキンライス」等の語が抽出されており地域 B の代表例である．地域 B は，R 群，S 群の語の割合が大きい地域である．地域 A と同様に考えると，抽出される語に普通名詞が多いことを示し，居住区のような，比較的一般性が高いと予想される地域が多く含まれる．

地域特徴語群の重要度の遷移とその遷移により分類される語群の割合を空間セルの特徴として，空間セル同士の類似度をはかることで，新しい地域分類ができると考えられる．語の重要度に基づき，地域特徴語が一致する数の多さで地域の類似度をはかる従来の手法では，空間的自己相関により近い地域のみが分類されることが多いが，地域特徴語群の重要度の遷移による分類手法では，離れているセルであっても，似た地域として分類できる点が大きな利点である．

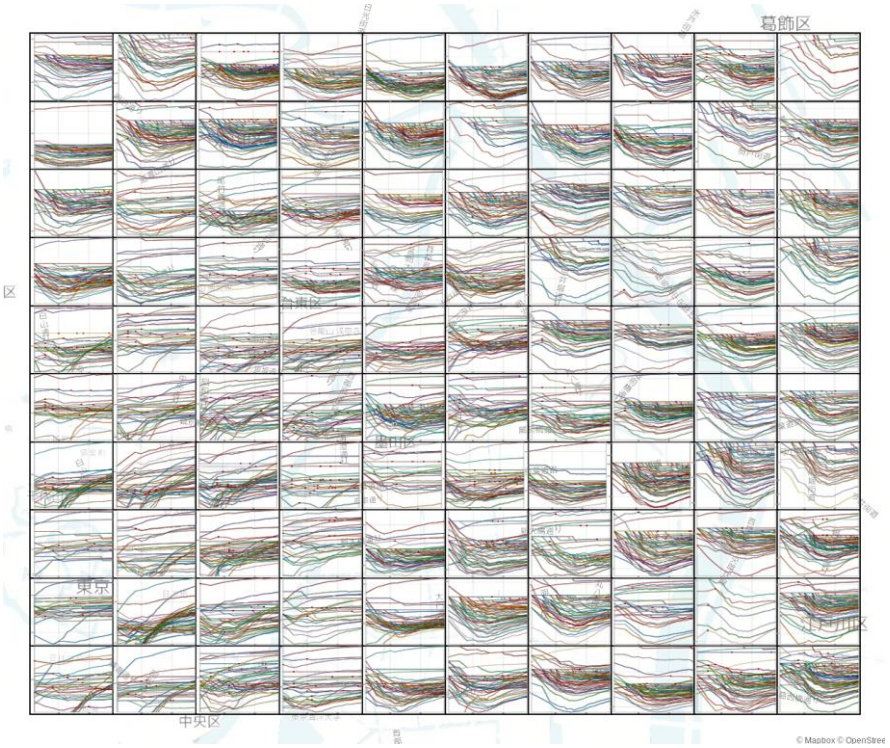

図 3 各空間セルにおける地域特徴語群の重要度の遷移

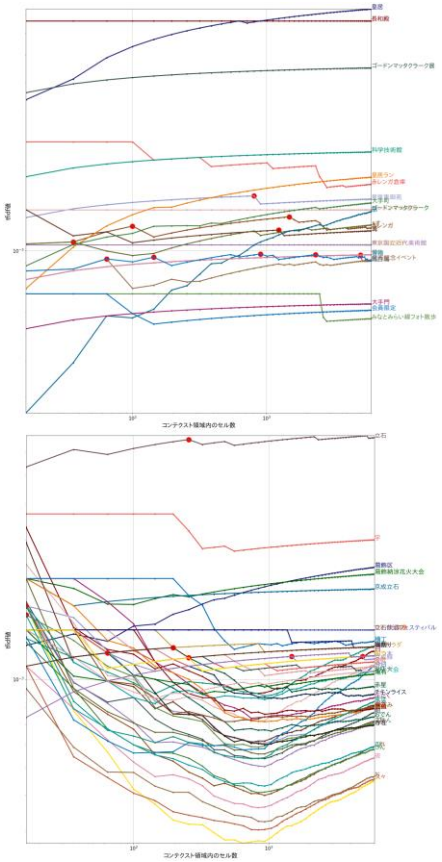

図 4 地域 A(上段)と地域 B(下段)の地域特徴語群の重要度の遷移例

7. まとめ

本研究では，スケールの変化を考慮した語の重要度を定義した．その後スケール変化に応じた重要度の遷移に基づき地域特徴語を分類し，分類の妥当性を評価し，地域特徴語として各語がもっとも重要となるスケールが定まることを示した．また，地域特徴語群の重要度の遷移によって地域を分類することができる可能性を示した．今後は，コンテクスト領域を変化させるだけでなく，空間セルの大きさを変化させることで，面積単位地区問題をある程度考慮した地域特徴語の抽出を目指す．また，地域特徴語群の重要度の遷移のクラスタリングに基づく地域分類の手法を検討する．

参考文献

[1] Mehta, P. et al., μTOP: Spatio-Temporal Detection and Summarization of Locally Trending Topics in Microblog Posts, In Proc. of EDBT, pp.558-561, 2017.
[2] Feick R. et al., Identifying Locally- and Globally-Distinctive Urban Place Descriptors from Heterogeneous User-Generated Content, Advances in Spatial Data Handling and Analysis, Springer, pp.51-63, 2015.
[3]鈴木英之,2017,位置情報付きソーシャルメディア等を用いた観光地域ブランドの評価,地理情報システム学会研究発表大会講演論文集