6P-04

語の一般性と空間領域との関係に着目した位置情報付き SNS からの地域特徴語抽出

秋庭　武†　　藤田　秀之†　大森　匡†　　新谷　隆彦†
電気通信大学情報理工学部情報・通信工学科†

1. はじめに

Twitterを代表とするSNSへ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，基本的な要件である．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 tf-idf 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．

本研究の対象は，位置情報付きツイート(Twitter の投稿データ)pの集合である．位置情報付きツイート$p =< loc, text >$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．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．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，グリッドのセル(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(語の頻度付き集合)を集約し，ひとつの文書とみなす．ここで文書とは，語の頻度付き集合である．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 tf-idf 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．

tf-idf 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．全国スケールと都市スケールの 2 種類のコンテクスト領域による地域特徴語の比較[2]が報告されている．本研究では，コンテクスト領域を多段階に変化させて，地域特徴語と重要度の変化を分析する．

Extraction of location-specific terms from location based SNS focusing on term generality and geospatial region
Takeru AKIBA, Hideyuki FUJITA, Tadashi OHMORI, Takahiko Shintani
Department of Information Science and Engineering, The University of Electro-Communications

2. 提案手法

まず，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，提案手法で用いる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．ビュー領域$V \subseteq D$は，可視化の対象となる画面表示領域である．ユーザの地図操作(ズームやパン)等により指定される．コンテクスト領域C($V \subseteq C \subseteq D$)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．

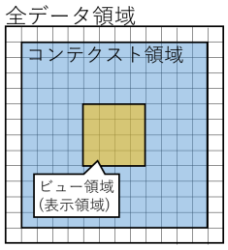

図 1. 領域定義

提案手法は，初等的手法における tf-idf 値の算出方法を拡張した手法である．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，次で定義する．

$$d_{w,i,C} = tf_{w,i} * idf_{w,C}$$

ここで，$tf_{w,i}$は空間セルiにおける単語wの出現頻度を，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．

$$tf_{w,i} = \frac{fr_{w,i}}{\sum_{v \in i} fr_{v,i}}$$

ここで$fr_{w_i,i}$は空間セルiにおける語$w_i \in i$の出現頻度である．次

に，$idf_{w,C}$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り，以下のように定義する．

$$idf_{w,C} = \log(\frac{|C|}{df_{w,C}})$$

ここで，$df_{w,C}$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 tf-idf 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，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(一般性がもっとも低い)語となる．

$$d_{base_i,i,C} = a_i * \log(|C|)$$

となる．ここでa_iはi内での$base_i$の出現頻度を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，a_iは定数であり，コンテクスト領域Cが変化しても変化しない．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．

$$d'_{w,i,C} = \frac{d_{w,i,C}}{d_{base_i,i,C}}$$

3. 実験

地域を分割するグリッドとして，総務省の定める標準地域メッシュのうち3次メッシュを用いる．各空間セルは約 1km 四方の矩形となる．データとして，2018年7月1日から2018年9月1日までに集計された約 340 万件の位置情報付きツイートを用いる．ビュー領域を上野公園周辺の 4 セル(メッシュコード 53394651, 53394652, 53394661, 53394662)とし，それぞれセル1から4とする．

コンテクスト領域の大きさ(スケール)を，ビュー領域から全データ領域へと拡大させ，コンテクスト領域の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．具体的な拡大手順として，コンテクスト領域の一辺の空間セル数を2^n (n = 1,2, ...)とし，各コンテクスト領域の中心を，ビュー領域の中心とする．ここで，コンテクスト領域にデータが存在しないセルが含まれる場合，そのセルをコンテクスト領域から除く．

以降のグラフの横軸は，全てコンテクスト領域内の空間セル数(スケール)である．また，ここでは例としてセル 1 に関するグラフを示す．　縦軸を各スケールにおける tf-idf 値とした地域特徴語群の重要度の遷移を図3に，基準語により正規化した tf-idf 値とした重要度の遷移を図4に示す．図3では，最広域に向けて多くの語が収束していくのに対して，図4では，最広域でも語が比較的分散しており，また，後述するピークも出現し，基準語による正規化の効果として，スケール変化に伴う各語の重要度の遷移と，語ごとの差異がより明確に示されている．また，図4では，各語の重要度が大きくなることは，各スケールにおいて最も重要度が高い基準語の tf-idf 値に，各語の tf-idf 値が近付くことを示す．よって，重要度が右肩上がりの部分は，その語がそのスケール範囲で一般的でなくなる(固有性をもつ)ことを示し，重要度が右肩下がりの部分は，その語がそのスケール範囲で一般的になることを示す．すなわち各語が最も特徴を持つスケール範囲は，重要度のピーク(極大値)であることが示せる．例えば，「野外(野外ステージ)」はコンテクスト領域内セル数が x=4 から x=16 で重要度の極大値をとっている．これは「野外(野外ステージ)」がコンテクスト領域セル数が4から16のスケールで固有名詞的に使われていることを示す．また，スケールがさらに拡大されると，「野外(野外ステージ)」の重要度は右肩下がりとなる．これは本語が各スケールで固有の意味を持たなくなることを示す．この場合では，コンテクスト領域セル数が16より増えセル1以外の場所に「野外(野外ステージ)」を表す場所が含まれたことを示す．一方で，市区町村名(「台東」「上野」「東京」)のピークはグラフ中に存在していない．これは，これらの語が固有名詞だからである．すなわち，どのスケールでも一つの意味を表す語のピークは存在しない．「上野公園」「アメ横」「アメヤ」のピークがグラフ中に存在しないのはこのためである．

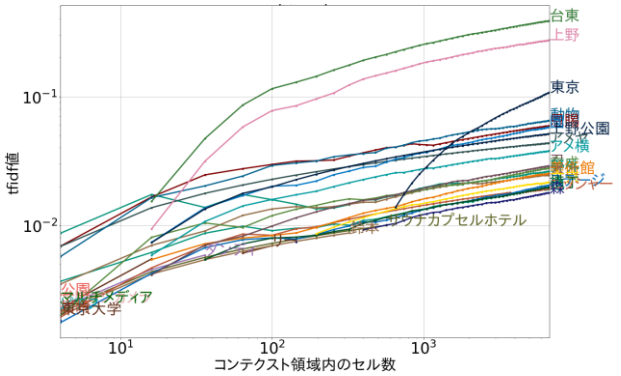

図 3. 地域特徴語群の tf-idf 値の遷移

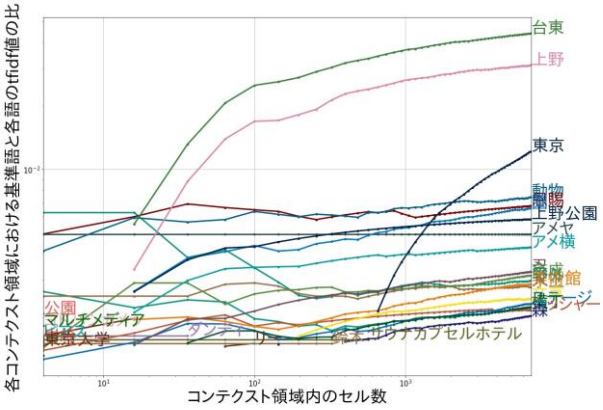

図 4. 地域特徴語群の正規化 tf-idf 値の遷移

4. おわりに

重要度として tf-idf 値を用いた地域特徴語の抽出において，idf の算出に用いる地理的領域の広さを多段階に変化させて，語の重要度の変化を分析し，スケールを考慮した地域特徴語の抽出と可視化に向け，有用な性質を確認した．

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, pp.51-63, 2015.