ジオタグ付きツイートで用いられる名詞の空間的広がりと階層性
－京都市における事例分析－

桐村　喬・藤原直哉・平岡喬之

Spatial Distributions and Hierarchical Structures of Nouns Used in Geo-tagged Tweets: A Case Study in Kyoto City

Takashi KIRIMURA, Naoya FUJIWARA, and Takayuki HIRAOKA

Abstract: We study spatial distributions of nouns used in geo-tagged tweets in Kyoto city in order to understand characteristics of the nouns tweeted in specific locations inside the city. Here, TF-IDF is employed to quantify importance of nouns in each location. We found that there exist place names, area names, and facility names, whose TF-IDF values are heterogeneously distributed in the city, resulting in non-trivial spatially connected components. These connected components reflect our perception about geographical regions indicated by these names.

Keywords: 形態素解析（morphological analysis），TF-IDF，Twitter，地名（place name）

1. はじめに

ジオタグ付きのツイッターの投稿（ツイート）データには，位置情報とツイート本文が含まれており，それをもとにして様々な空間分析が行われている（桐村，2019）．多くの場合，ジオタグを意識的に付与する場合のツイートの本文は，特定の場所を知らせるための内容となり，ツイート内容には地名をはじめとする様々な単語が含まれることになる．一方で，無意識的にジオタグが付与されている場合，独り言に近い“つぶやき”や他のユーザーとの何気ない会話も含まれ，より多様な単語が含まれることになる．

ジオタグ付きツイートに含まれる単語は，その地域のイメージやユーザーの空間認識と関連している．例えば，「京都」という地名であれば，それぞれのユーザーが京都であると考える地域において，「今，京都に着いた」などのツイートをすることになる．反対に，特定の地域におけるツイートに含まれる単語に注目すれば，ユーザーがもつ特定の地名などに関する空間認識の具体的な範囲を把握することができる．

一方で，地名以外にも，地名のような役割を果たす単語も存在する．例えば，都市のなかに 1 か所しかないような特定の商業施設名であれば，その施設名は，その都市においては地名と同等の働きをなすことになる．したがって，地名だけでなく，様々な単語，特に名詞に注目する必要がある．

そこで，本研究では，ジオタグ付きツイートの投稿本文で用いられる名詞に着目して，各名詞が特徴的に出現する空間的広がりと空間的スケールの階層性との関係を分析し，いくつかの名詞の基本的な地理的特性について整理する．

従来，アンケート調査結果などに依らない，網羅的な特定の GIS データを用いた地域イメージの研究には，建物名称が主に用いられてきた（小

桐村　喬
皇學館大学文学部コミュニケーション学科
E-mail: t-kirimura@kogakkan-u.ac.jp

池ほか, 2019)．また，画像投稿・共有サービスである Flickr のデータを利用したイメージの研究も行われているが（末田ほか, 2011)，対象地域は限定的である．本研究で用いるジオタグ付きツイートについては，近年，ポイント単位でのデータが少なくなっている問題も指摘されているが（桐村, 2019)，それでも，網羅的に都市単位での分析ができる資料として一定の価値を有している．

2. データと分析方法

2.1 分析の手順

分析の対象地域は京都市を含む矩形の範囲であり，ツイッターの Public streams API を利用して取得された，この地域における 2014 年 1 月から 2018 年 3 月までの約 740 万件のポイント単位のジオタグ付きツイートデータを利用する．そして，ジオタグ付きツイートデータの本文の情報について，NEologd の辞書を内包した Python 用の形態素解析ライブラリである Janome を利用して形態素解析を行う．形態素解析の結果として抽出された一般名詞と固有名詞（以下，合わせて名詞と総称する）について，正方形のグリッドごとに，後述する空間的 TF-IDF を算出する．グリッドの 1 辺の長さは，複数の空間的スケールから分析するために，50m, 100m, 200m, 400m, 800m, 1,600m, 3,200m の 7 種類とした．

TF-IDF は，一般的には，文書内での全単語に占める特定単語の出現頻度（Term Frequency, *tf*）と，文書単位でみたときの全文書における特定単語の出現文書の比率の逆数（Inverse Document Frequency）の対数（*idf*）が用いられる．通常は文書単位で算出されるものの，ツイートデータの場合，1 つのツイートを 1 文書としてしまうと，特定の単語が繰り返し 1 つのツイートで用いられることはまれであり，ほとんどの場合，*tf* は低い値になってしまう．ここでは，ツイートデータに合わせた空間的 TF-IDF を定義し，それぞれの名詞について値を求めて，空間的スケールごとの特徴について整理する．

2.3 空間的 TF-IDF の算出

空間的 TF-IDF は，一般的な TF-IDF が文書単位で定義されるのに対し，各グリッドをひとまとまりとして定義される点に特徴がある．すなわち，特定のグリッドにおける複数のツイートはまとまった 1 つの文書として扱われることになる．具体的には，以下のような式で定義される．

$$tf_{ij} = \frac{n_{ij}}{N_j} \quad \cdots\cdots (1)$$

$$idf_i = \log_2 \left(\frac{C}{c_i} + 1 \right) \cdots\cdots (2)$$

$$tfidf_{ij} = tf_{ij} \cdot idf_i \quad \cdots\cdots (3)$$

(1) 式は，グリッド j における名詞 i の使用頻度である．ここで，$n_{i,j}$ は，グリッド j における名詞 i を使用したユーザー数であり，N_j は，グリッド j における総ユーザー数である．*tf* の計算の基準を名詞数ではなく，ユーザー数としたのは，特定のグリッドでより多くのユーザーに使われている名詞ほど，重要度が高いと考えるためである．(2) 式は，名詞 i に関する対象地域内での重要度を求めるものである．C は，対象地域内におけるグリッドの総数を示し，c_i は，名詞 i が使われているグリッドの数を示している．(3) 式は，グリッド j における名詞 i の空間的 TF-IDF を求めるものであり，*tf* と *idf* を乗じることで求められる．

特定のグリッドにおいて，特定の名詞の空間的 TF-IDF が高ければ，その地域で，その名詞を使用するユーザーが多いことになり，その地域を代表する名詞と考えることができる．

一方で，N_j が小さい地域においては，*tf* が過大になりやすく，それほど重要な名詞ではないと考えられる場合でも，空間的 TF-IDF が大きな値になってしまうことがある．このような外れ値の影響を軽減するために，カーネル密度推定法を用いて空間的 TF-IDF の平滑化を行う．カーネルの半径はグリッドサイズの 2 倍とした．

名詞ごとに，このような処理を施したうえで，空間的 TF-IDF があるしきい値以上を示す範囲を特定しディゾルブして，グリッドが空間的に連坦（連結）する最大の領域（最大連結成分）を求める．各名詞の最大連結成分は，その名詞の重要度が高い地域を指しており，ツイッターユーザーの特定の名詞に対して抱く空間的なイメージの具体的な領域と考えることができる．空間的 TF-IDF のしきい値については，本研究では 0.2 とした．

3. 分析結果と考察

3.1 ユーザー数上位 5 語の特徴

形態素解析によって，一般名詞または固有名詞と判断されたもののうち，ユーザー数上位 10 語をみると，顔文字などでよく用いられる「o」(4 位）や「ω」（7 位）がみられ，顔文字の形態素解析が十分ではない可能性が指摘できる．また，「ほんま」（9 位）は副詞的な意味を持つ方言であり，名詞ではなく，「きた」（10 位）も多くは「来た」と考えられる．このような問題に対しては，ツイッターでの特有の表現に合わせて，形態素解析の方法を改善していく必要があるが，ここでは，名詞と考えて差し支えないものについての上位 5 語（「京都」（1 位），「人」（2 位），「場所」（3 位），「写真」（5 位），「自分」（6 位））に注目する．

図-1 は，これらの 5 語に関するグリッドサイズと最大連結成分の面積の関係を示したものである．いずれの語についても，グリッドサイズが大きくなるほど，最大連結成分の面積も大きくなる傾向にある．ただし，「人」については，100m で面積が大きくなる傾向が確認でき，他の 4 語とは様相がやや異なる．

最も小さいグリッドである 50m グリッドに注目して，5 語の最大連結成分の分布を示した（図-2）．最も面積が広いのは「人」であり，市街化された盆地の大部分を占めているが，西部の嵐山までは及んでいない．「京都」は「人」よりも範囲が狭く，中心部の市街地とその周辺に，道路に沿って広がっている．「京都」は嵐山だけでなく，南部の伏見にも到達していない．「場所」，「写真」，「自分」については，中心部とその周辺の観光地（東山）に広がっており，これらの語と，観光や買い物行動などとの関係の深さがうかがえる．

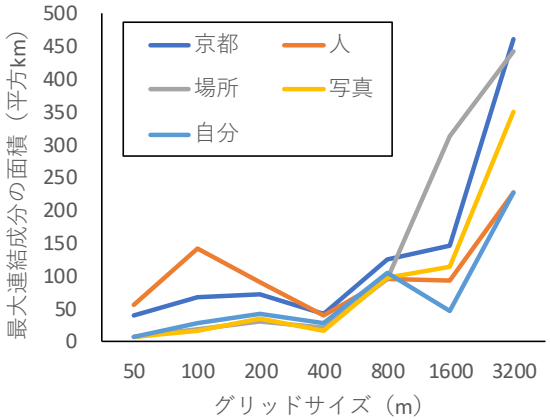

図-1　グリッドサイズと最大連結成分の面積

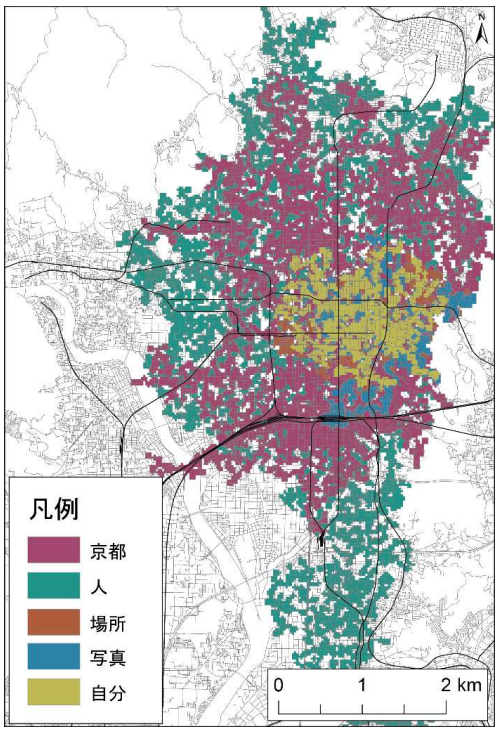

図-2　最大連結成分の分布
（京都/人/場所/写真/自分・50m）
※面積が大きいほど背面に表示している．

3.2 地名に関する最大連結成分の分布

「京都」を除けば，京都における特定の地域の名称は，それほど頻繁には使用されていない。ま

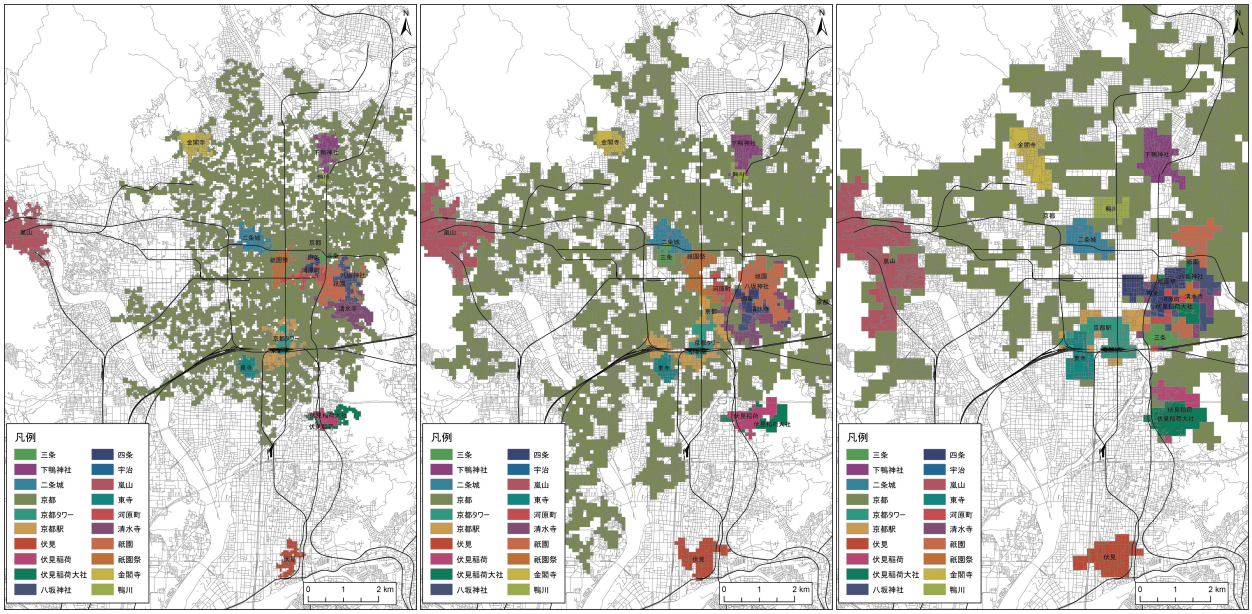

図-3　主な地名の最大連結成分の分布（左：50m/中：100m/右：200m）

※面積が大きいほど背面に表示している．

た，形態素解析では，地名であっても一般名詞と判断されているものもある．これらの名詞から，地名と考えられるものをすべて抽出することは難しい．そこで，地区名や通り名，寺社名，施設名など，地名またはそれに類するものと判断できる名詞のうち，使用ユーザー数上位 20 位の名詞の最大連結成分について，50m，100m，200m のグリッドのものを示した（図-3）．なお，複数の品詞が存在する単語は，よりユーザー数が多いものに限定した．

「京都」は，グリッドサイズによって分布範囲が大きく変動し，50m と 200m は対照的な分布となっている．「嵐山」や「二条城」，「下鴨神社」，「祇園」，「東寺」，「伏見」などは，3 つのグリッドサイズのいずれでも，面積および場所がある程度安定している．しかし，「金閣寺」や「祇園」は 200m では，その範囲を大きく拡大させており，実際の地域との乖離が大きい．「祇園祭」は地名ではないものの，50m と 100m では，山鉾町を中心とする地域に最大連結成分がみられる．しかし，200m では，山鉾町から大きく離れている．

これらの地名に関する最大連結成分の分布範囲を確認する限り，200m よりも大きなグリッドサイズでは，分布範囲が本来の場所と乖離しやすい傾向にあると考えられる．カーネル密度推定による平滑化の影響もあり，大きなグリッドほど平滑化されることで，かえって本来の分布範囲よりも大きくなることになる．空間的 TF-IDF およびカーネル密度推定による平滑化による本手法を用いて，それぞれの地名の空間的イメージを捉えようとする場合は，少なくとも京都においては 100m 以内の大きさのグリッドが適していると考えることができる．しかし，他の都市においては最適なグリッドサイズが異なると考えられ，いくつかの事例分析を通じた指標の開発が望まれる．

参考文献

桐村　喬編, 2019.「ツイッターの空間分析」，古今書院（出版予定）．

小池束紗，貞広幸雄，對間昌宏, 2019. 東京都区部における建物名称に用いられる地名の滲出現象，GIS－理論と応用，27(1)，25-31.

末田　航・味八木　崇・暦本純一, 2011. 実世界集合知による利用者の認知地図の可視化とモバイルインタラクションへの適用，情報処理学会論文誌，52(4)，1465-1474.