位置属性を有しない事物に対する地理的特定性の分析

陰山宗一 [1]　乾孝司 [1]
[1] 筑波大学大学院 システム情報工学研究群 情報理工学位プログラム
s2120577@s.tsukuba.ac.jp　inui@cs.tsukuba.ac.jp

概要

ある事物に対する言及が文書内に現れた時にその言及が地理的位置を特定する程度をあらわす指標として地理的特定性指標がある [5]。本稿では、地理的特定性指標の利用可能範囲を調査する一環として、地理的特定性指標の構成要素のひとつである名称専有性指標に注目し、地理的位置属性を有しない５つのカテゴリ（列車、特産品、祭り、苗字、植物）に属するエンティティへの言及を対象にして、名称専有性の値の特徴を分析したので、その結果を報告する。

1　はじめに

SNS サービスに投稿される文書にはジオタグのような地理的位置を示すメタ情報が付与される場合がある。近年、この位置情報を利用することで投稿文書データから「今どこで何が起きているか」を把握するソーシャルセンシング技術への関心が高まっているが、SNS 投稿への位置メタ情報の付与は縮小傾向にある [1]。このため、位置情報の補完処理として、位置情報付きでない投稿文書の投稿場所（地理的位置）を自動推定する文書ジオロケーションに関する研究が進められている [2, 3, 4]。

文書ジオロケーションでは、文書内に現れる地名やランドマークへの言及（例えば、「東京」や「東京タワー」）が、投稿がなされた地理的位置を推定する有力な手がかりになることが多いが、それらの中には地理的な曖昧性を含む場合もあり、文書ジオロケーションの誤り原因となる問題が存在する。例えば、「中華街」を含んでいる投稿は、その投稿がなされた地理的位置を、横浜や神戸などの中華街が所在する場所へ絞り込む手がかりになり得るが、特定の一箇所に絞り込むまでには至らない。この手がかりの地理的曖昧性問題へ対応するため、陰山ら [5] は、文書内に出現する、地名やランドマーク等の地理的位置属性をもつエンティティへの言及に対して、その地理的位置の特定のしやすさを表す指標である地理的特定性指標を提案し、Wikipedia データの情報に基づいて具体的な指標値を求めた。そして、地理的特定性指標の情報を文書ジオロケーションへ適用することで、その有効性を報告している。しかし、彼らの報告では、地理的特定性指標の文書ジオロケーション以外への活用に関しては議論がなされておらず、その活用可能範囲は未知である。

そこで我々は、陰山らの地理的特定性指標 [5] の活用可能範囲を明らかにすることを目標に研究を進めているが、その一環として現在、各対象に対して推定した指標値がどのような特徴をもつかを分析している。具体的には、文献 [5] では、文書ジオロケーションへの適用を見据えていたため、地理的位置属性をもつエンティティへの言及のみに注目して指標値を推定していたが、本研究では、地理的位置属性をもたないエンティティへの言及に対して指標値を求め、指標値の観察を通して、その特徴分析を試みる。地理的位置属性をもつエンティティの場合、基本的に、地理的位置属性はエンティティの所在地をあらわし、地理的特定性の値へも所在地情報が反映される。しかし、地理的位置属性をもたないエンティティの場合は所在地という概念が無い（あるいはあっても希薄である）ため、地理的特定性の値が何を反映するかは自明ではない。

なお、地理的特定性指標は、地理的曖昧性と名称専有性の対で構成されるが、このうち地理的曖昧性は同一の名称をもつ対象の数を表し、直感的にもわかりやすい。そこで今回は特に名称専有性に注目し、その分析結果を報告する。

2　名称専有性

名称専有性（name exclusivity）は、エンティティと言及との対応関係に関する人々の一般的な認知の度合いに関する指標である。例えば、一般的な文脈において現れた「横浜」という言及は、高知県や福岡県に所在する地域である横浜よりも人々の認知度のより高い地域であると考えられる神奈川県横浜市の事であると認知されやすいだろう。このように、名称が同じエン

ティティが複数存在したとしても、それらのエンティティがすべて同程度に認知されるとは限らず、「横浜」における神奈川県横浜市のように特定のエンティティに偏って認知される場合がある。名称専有性は、このような現象を捉えることを目的とした指標であり、ある対象が特定のものとして認知される程度をあらわす。

以下では、文献 [5] に従い、言及に対する名称専有性の値の求め方について述べる。この値は、エンティティに対する名称専有性の値から求めるため、まず、エンティティに対する名称専有性の値の推定方法について説明する。なお、以降の説明では、日本語 Wikipedia のエントリページをエンティティ、また、日本語 Wikipedia 内で各エンティティを参照する（アンカーリンクを張る）ために使用されているアンカ文字列をエンティティに対する言及であるとする。

日本語 Wikipedia において、地理的特定性を求めたいエンティティを p_t としたとき、この p_t へのすべての言及からなる言及集合を M_e とし、その要素を $m_i (i = 1, 2, ..., |M_e|)$ とする。次に、M_e の要素である各言及に対して、それがリンクしているすべてのエンティティを抽出する。抽出されたエンティティ集合を P_e とし、その要素を $p_j (j = 1, 2, ...|P_e|)$ とする。そして、ある言及 m_i が、あるエンティティ p_j へリンクを形成している回数を $l_{p_j}^{m_i}$ としたとき、このリンク回数がある対象が認知される程度を近似的に表していると考え、エンティティ p_t の名称専有性 $exc(p_t)$ は式 (1) で表されるとする。

$$exc(p_t) = \frac{\sum_{m \in M_e} l_{p_t}^{m}}{\sum_{p \in P_e} \sum_{m \in M_e} l_{p}^{m}} \quad (1)$$

ここで、$0 < exc(p_t) \leq 1$ である。

次に、$exc(p_t)$ を利用して、言及に対する名称専有性の値を求める。文献 [5] では、47 都道府県に分類する文書ジオロケーションへの適用が想定されていたため、言及に対する値は、エンティティのようなスカラー値ではなく、47 次元ベクトルである。名称専有性を求めたい言及を m_t としたとき、日本語 Wikipedia において、m_t がアンカ文字列となってリンクしているすべてのエンティティを抽出し、その集合を P_m とする。この時、Okajima ら [6] の手法に基づき、P_m の各要素となるエンティティ p_j に対して、そのページの本文内容に初めて登場する都道府県名の情報をそのページの都道府県要素として記憶する。次に、この都道府県要素の値に基づいて、P_m を 47 個の部分集合に分割する。ある都道府県 k に対応する P_m の部分集合を P_m^k としたとき、P_m^k の要素に従って、都道府県 k に関する、エンティティの名称専有性の最大値 $max\_exc(k)$ を求める。

表 1　言及「厳島神社」に対する名称専有性の値

北海道	0.015
...	...
静岡県	0.021
...	...
京都府	0.019
...	..
兵庫県	0.014
...	...
広島県	**0.646**
...	...
沖縄県	0

$$max\_exc(k) = \begin{cases} \max_{p \in P_m^k} exc(p) & P_m^k \neq \phi \\ 0 & otherwise \end{cases} \quad (2)$$

最後に、以上で求めた 47 都道府県に対応する $max\_exc(k)$ を要素とする 47 次元ベクトルを構成し、これを言及 m_t に対する名称専有性の値とする。

例として、後述の分析作業と同じデータを用いて言及「厳島神社」に対して求めた名称専有性の値を表 1 に示す。日本国内には「厳島神社」という名称をもつエンティティは複数存在するが、広島県廿日市市に所在するものが一般的な認知度がもっとも高いと予想される。表の値は確かに広島県の値が高くなっていることが確認できる。なお、この例でも示されているように、名称専有性の値は形式的には 47 次元を仮定しているが、ベクトル要素の値が 0 となる次元（都道府県）も存在する。

3　名称専有性の分析

3.1　分析対象

本研究では、住所をもたないエンティティを地理的位置属性をもたないエンティティとして、以下のカテゴリに属するエンティティを選択し、それらへの言及に対する名称専有性を分析対象とした。

- 【列車】
- 【特産品】
- 【祭り】

- 【苗字】
- 【植物】

このうち、【列車】および【特産品】カテゴリのエンティティは、直接的には住所をもたない[1)]が、出発地や産地として間接的に特定の地域と関連が深いと考えられる。それら関連地域での名称専有性の値がどのような値かをこの後確認していく。つぎに、【祭り】のエンティティは、実施される地域が必ず存在するため、実施地域を地理的位置属性としてもつエンティティであるとも言えるが、ランドマーク「東京タワー」のようなモノ的なエンティティではなくコト的なエンティティであるため、その特徴を観察するために分析対象に含めることにした。【苗字】は、これまで述べたカテゴリに比べてエンティティ数が多いと考えられ、そのような場合の特徴を観察するために含めることにした。最後の【植物】は、これまで述べたカテゴリのエンティティは固有名を持ちやすい特徴があるため、そうでない場合の特徴を観察するために含めた。

3.2 データ

分析のために指標値の推定に用いる Wikipedia データは、文献 [5] と同じであり、2021 年 5 月 20 日付の日本語 Wikipedia ダンプデータ[2)]を用いた。

3.3 分析結果

3.3.1 表記について

求めた名称専有性の値を確認する前に、以降で事例を示す際の表記について説明する。各言及に対する名称専有性を求めた結果、多くの事例で 47 次元のうちの多数の次元で要素の値が 0 となっていた。そこで、以下では要素の値が 0 以外となっている場合のみ記載する。要素の値が 0 以外となる次元が複数存在する場合は、値の降順に並べて示す。各言及事例において、言及の一部に地名を含む場合はその箇所を下線で示す。また、推定結果が誤りと思われる箇所は丸括弧をつけて示す。

前節で述べたように、名称専有性の値は、ベクトルの各要素の最大値が 1 であり、要素和の最大値が 1 ではない。以降の分析結果を読む際は注意されたい。

1) 例えば、駅舎は住所をもつが、列車の車両自体は住所をもたない。
2) https://dumps.wikimedia.org/jawiki/ から取得。

3.3.2 カテゴリ 【列車】

列車のうち、急行列車や特急列車などの列車名をもつエンティティへの言及に対して名称専有性の値を求めた。その結果、走行路線上の地域で名称専有性の値を持ちやすいことが確認できた。走行路線によっては都道府県をまたぐ場合があるため、複数の地域で高い名称専有性を有する可能性があるが、走行路線上の地域を網羅することはなく、網羅性に関しては不十分であることがわかった。なお、網羅性に関しては、以降のどのカテゴリでも同様な傾向であった。

- 踊り子　静岡県:1
- きたぐに　福井県:1
- ムーンライト信州　山梨県:1
- 阿波　徳島県:0.974, 岡山県:0.119,（千葉県:0.016）
- うずしお　徳島県:0.908, 愛媛県:0.685

3.3.3 カテゴリ 【特産品】

ご当地グルメを含む各地域の特産品を対象に名称専有性の値を確認する。地域特産品のうち、名称に地名を冠しているエンティティへの言及では、そのエンティティを生産したり、水揚げする地域が名称専有性の値を持ちやすく、かつ、それらは名称に冠している地名が所在する都道府県と一致することが多かった。「讃岐うどん」は、都道府県の中では香川県で最大値をもっていたが、その値は小さい。この原因を調べたところ、広く普及し知名度が高いと考えられるエンティティは Wikipedia 内での言及回数が多く、それに合わせて言及表現が多様となる場合があり、特に地名（「讃岐」）や、「うどん」のような当該エンティティの上位概念となるエンティティをあらわす表現で言及されると、それらとの間で指標値が分散されるため、値が低くなりやすいことが確認できた。

- 喜多方ラーメンバーガー　福島県:1
- 水沢うどん　群馬県:1
- 川俣シャモ　福島県:1
- 関あじ　大分県:1
- 讃岐うどん　香川県:0.127

つぎに、名称に地名を冠していないエンティティへの言及では、当然ながら、エンティティの名称には地名を連想させる手がかりが含まれないものの、宮城県仙台市で有名な「牛タン」など、そのエンティティで知られている地域が値を持つことができている。ただし、名称に地名を冠しているエンティティへの言及と

は異なり、一般的な名称としての言及（「牛タン」の場合は特産品ではなく食品としての対象）との間で表現の重なりが生じやすいため、名称専有性の値は高くなりにくい特徴がある。

- 牛タン　宮城県:0.522
- スタミナラーメン　茨城県:0.833, 埼玉県:0.429
- 玉子焼　兵庫県:0.741

3.3.4　カテゴリ【祭り】

祭りは、各地域で実施される行事であるため、その実施地域との関連が深いと考えられる。「さっぽろ雪まつり」や「仙台七夕まつり」といった名称に地名を冠する行事では、先述の地名を冠する地域特産品と同様に、名称に冠している地名に対して高い名称専有性の値を持ちやすいため、ここでは、名称に地名を冠しない場合について確認する。この場合の事例を見ると、それぞれの祭りの実施で有名な地域が名称専有性の値を持ちやすいことが確認できた。「さっぽろ雪まつり」ではなく「雪まつり」のように、実施地域をあらわすであろう地域を冠しないエンティティへの言及を分析対象としているにも関わらず、このような結果が得られた。このことから、名称専有性の観点から言えば、祭りは、ランドマークなどの地理的位置属性を持つエンティティと同様の特徴をもつカテゴリと見なして良さそうである。

- 雪まつり　北海道:0.879, 新潟県:0.467
- ねぶた祭り　青森県:0.448
- だんじり祭り　大阪府:0.324
- よさこい祭り　高知県:0.510
- 七夕まつり　愛知県:0.786, 富山県:0.714, 宮城県:0.046

3.3.5　カテゴリ【苗字】

苗字と地域の関連性としては、その苗字の起源となる地域や、その苗字を持つ人の人口が多い地域などが考えられる。各苗字に対する推定値を確認した結果、苗字の起源となる地域で値を持つ傾向があった。一部の苗字では、人口が多い地域と名称専有性の値に相関が見受けられたが、多くの苗字に関してはそのような特徴は見受けられなかった。このカテゴリのエンティティは、「大井」のように、要素が 0 以外の値をもつ次元の数が今回扱った他のカテゴリに比べて多い特徴があった。このカテゴリはエンティティ数が比較的多く、また、他カテゴリと多様な関係性をもつと考えられ、そのことが結果に影響していると考えられる。

- 河西　北海道:1
- 川名　神奈川県:1
- 米原　滋賀県:0.913
- 芳賀　栃木県:0.994
- 磯部　山口県:0.671, 群馬県:0.575, 三重県:0.531, 茨城県:0.176
- 大井　東京都:0.844, 静岡県:0.536, 三重県:0.25, 埼玉県:0.084, 岐阜県:0.0831, 神奈川県:0.0243, 山梨県:0.0153

3.3.6　カテゴリ【植物】

今回、固有名でないエンティティへの言及に対する名称専有性の値を確認するために、その例として植物に注目した。その結果、まず、固有名でない場合でも固有名の場合と同じように名称専有性の値をもつことが確認できた。植物では、群生地をもつ都道府県が値を持ちやすいことがわかった。また、植物の中には、その果実や、植物から抽出した油などの関連商品を生産でき、それらの生産を目的とした植物栽培が盛んな地域で大きな値を持つ傾向が確認できた。

- エゾマツ　北海道:1
- サクラソウ　北海道:1
- ヒバ　青森県:1
- ウメ　和歌山県:0.939,（鹿児島県:0.086）
- オリーブ　香川県:0.914, 三重県:0.24, 東京都:0.085

4　まとめ

本稿では、地理的位置属性を有しないエンティティへの言及を対象にして、地理的特定性指標の構成要素のひとつである名称専有性の値を求め、その特徴を分析した。その結果、今回題材にした５つのカテゴリに関して幾つか特徴があることが確認できた。地理的特定性指標の活用可能範囲を明らかにするためには、今後も継続的な分析が必要である。例えば、本稿で示した分析結果は、一部のエンティティに対する限定的な結果であり、規模を広げた詳細かつ定量的な分析を実施する必要がある。また、名称専有性の値は元データである Wikipedeia のページ記述スタイルの影響を受けるが、エンティティのカテゴリごとに影響の度合いを調査する必要がある。

謝辞

本研究の一部は JSPS 科研費 21K12137 の助成を受けたものです。

参考文献

[1] Anna Kruspe, Matthias Häberle, Eike J. Hoffmann, Samyo Rode-Hasinger, Karam Abdulahhad, and Xiao Xiang Zhu. Changes in Twitter geolocations: Insights and suggestions for future usage. In **Proceedings of the Seventh Workshop on Noisy User-generated Text**, pp. 212–221, 2021.

[2] Bo Han, Paul Cook, and Timothy Baldwin. Geolocation prediction in social media data by finding location indicative words. In **Proceedings of International Conference on Computational Linguistics**, pp. 1045–1062, 2012.

[3] Lianhua Chi, Kwan Hui Lim, Nebula Alam, and Christopher J Butler. Geolocation prediction in Twitter using location indicative words and textual features. In **Proceedings of the 2nd Workshop on Noisy User-generated Text**, pp. 227–234, 2016.

[4] Jey Han Lau, Lianhua Chi, Khoi-Nguyen Tran, and Trevor Cohn. End-to-end network for Twitter geolocation prediction and hashing. In **Proceedings of the Eighth International Joint Conference on Natural Language Processing**, pp. 744–753, 2017.

[5] 陰山宗一, 乾孝司. 言及に対する地理的特定性指標の提案と文書ジオロケーションへの適用. 情報処理学会自然言語処理研究会 (NL-253-19), 2022.

[6] Seiji Okajima and Tomoya Iwakura. Japanese place name disambiguation based on automatically generated training data. In **19th International Conference on Computational Linguistics and Intelligent Text Processing**, 2018.