Web文書を利用したPOIに関連する地名の抽出

今井 良太[1]　数原 良彦[1]　戸田 浩之[1]　鷲崎 誠司[1]

概要：
飲食店や観光スポットのような Point of Interest (POI) がどこにあるかを表す情報としては，住所や経緯度がよく用いられる．一方で，日常会話やブログ，SNS 等で POI の所在地を表すために用いられる場所の表現には，住所として存在しないものや，正式な区画と必ずしも一致しないものも多い．本研究では，POI の所在地を表すために用いられる場所の表現を POI の関連地名と呼び，これをブログ等の Web 上の文書から抽出することによって，POI に関連地名を付与する手法を提案する．実験では，POI に 5〜10 件の関連地名を付与する場合について提案手法を適用し，適したパラメータの検討を行った．

1. はじめに

飲食店や観光スポット，ランドマークのような，実世界における意味のある点は Point of Interest (POI) と呼ばれ，人々の移動の支援やマーケティングについて考える上で重要なものである．例えば，飲食店を様々な条件で検索して食事をする店を決めるためのサービスを提供したり，ある地域の観光スポットやランドマークの情報をまとめて訪問者数の向上をねらったりするために必要となる．

POI に付与される重要な情報として，位置情報がある．POI には, そのデータの利用形態や POI のジャンルによって様々な情報が付与される．例えば飲食店を案内するためのデータなら，各飲食店には営業時間や駐車場の有無が付与され，観光スポットのガイドのためのデータなら，各スポットには歴史や直近のイベント等が付与される．一方で，POI の所在地についての情報は利用形態やジャンルによらず必要な情報であるといえる．

POI の位置情報には住所や経緯度がよく用いられるが，これらは POI の大まかな所在地を把握する用途には向いていない．例えば，「山下公園」は住所としては「神奈川県横浜市中区山下町 279」にあり，経緯度では緯度=35.446，経度=139.650 付近である．「山下公園」の大まかな所在地としては，「関内」，「横浜港」，「横浜中華街」といった表現が考えられる．これらは観光ガイド等でエリアごとに情報をまとめる際にも用いられることがあり，有名であるといえるが，この周辺の土地勘がない人には，前述の住所や経緯度からこれらの表現を連想することは難しい．これら 3 つの表現はいずれも住所としては存在せず，一方で経緯度から周辺の地図を見ても，その場所がどう呼ばれているかを見つけることは難しいためである．

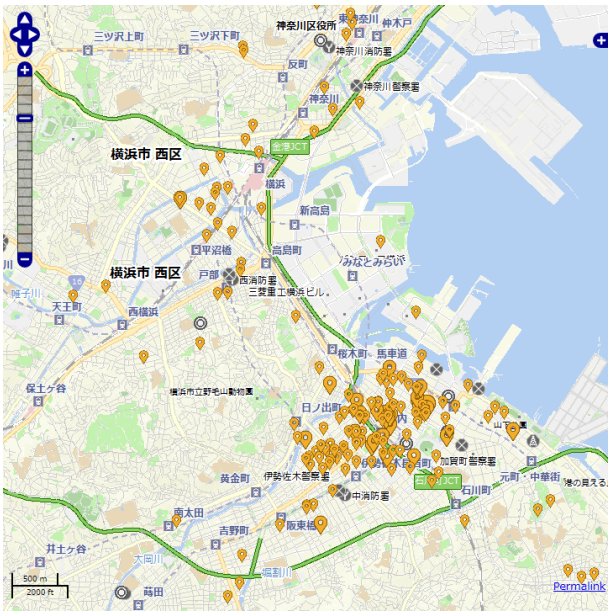
POI に対して，住所や経緯度に加えて関連する場所の表現を付与することができれば，POI の所在地を調べるときの利便性を向上させることができる．本研究では，この POI に関連する場所の表現を POI の「関連地名」と呼ぶ．

POI に関連地名を付与することは容易ではない．関連地名には「関内」のような住所として存在しないものが含まれるため，そのようなものを網羅する辞書を用意することは難しい．たとえ対応する住所があっても，関連地名から一般的に想起される範囲と住所上の区画は必ずしも一致しないため，そのまま用いることはできない．例えば，「品川」という関連地名に対しては「東京都品川区」という住所がある．一般的には品川駅周辺の POI も「品川にある POI」として想起されるが，「東京都品川区」には品川駅は含まれない．

本研究では，POI の集合が与えられたときに，Web 上の文書を用いて POI の関連地名を抽出する手法を提案する．Web 上の文書から POI の名称と文内共起する関連地名を抽出し，その共起回数に基づくスコアと，地理的な距離に基づくスコアを定義し，それらを用いて POI と関連地名の関連度を求める．そして，関連度の高い関連地名を選択することにより，有名なために共起回数が多くなるが地理的には離れている誤った地名が出力されることを防ぐ．

本研究で提案する，Web 上の文書内の文内共起回数に基づくスコアと地理的な距離に基づくスコアを同時に利用する手法の妥当性について述べる．5 節で述べる実験で利用した Web 文書群において，関連地名「関内」と文内共起する POI を地図上にプロットした例を**図 1** に示す．この例

[1] 日本電信電話株式会社 NTT サービスエボリューション研究所
NTT Service Evolution Laboratories, NTT Corporation



地図データ: Copyright(C)NTT 空間情報

図 1 関連地名「関内」に対応する POI

では，住所として存在しない関連地名「関内」に対応するエリアの付近に POI が密集していることがわかる．これにより，文内共起によって POI と関連地名を結びつけることが可能であることが示唆された．しかし，図 1 の例では関内駅周辺の POI に限らず，関内駅から離れた横浜駅周辺や，東神奈川駅周辺の POI も「関内」と共起してしまっていることがわかる．したがって，我々は共起回数のみではなく，地理的なスコアも同時に考慮する関連度を用いることで，これらの POI の集合の中でも，関内駅周辺の POI において「関内」を関連地名として優先的に選択することが可能になると考えた．

本稿の構成は次のとおりである．2 節では，本研究に関連する先行研究について述べる．3 節では，本研究で扱う問題を定義する．4 節では，本研究の提案手法について説明する．5 節では，提案手法の評価方法とその結果について述べる．最後に，6 節でまとめと今後の課題について述べる．

2. 関連研究

Web 上の文書から地理的な情報を収集する研究について述べる．

地名ではなく POI そのものを収集する技術として，相良ら [1] は，既知の店舗に対する評判情報とあわせて，未知の店舗情報を抽出する手法を提案している．この手法では，既知の店舗情報として電話帳を用いている．

文書中の地名を扱う際に，地名が指す実世界上の位置の候補がわかっているが，一意に特定できない場合がある．平野ら [2] は，ある地名について，文書中の他の地名との距離と，地名の有名度を組み合わせることで，高精度に地名を同定する手法を提案している．

元々地理的な属性をもたない情報に地理的な属性を与えるという観点では，オブジェクトレベルサーチを地域情報に適用した手塚ら [3] の研究がある．この研究では，Web 検索で入力されるキーワード型のクエリについて，その検索結果のスニペットに含まれる POI の位置情報を用いて地理的な関連性を求めている．本研究で扱う関連地名も，文書から抽出した段階では地理的な属性をもたないため，POI の位置の平均をとることで代表的な位置を推定している．一方，手塚らの研究においては地名に限らず任意のキーワードを対象とするため，キーワードが 2 つ以上の離れた地点と地理的な関連をもつことが考えられる．そのため，混合ガウス分布を用いているという点が本研究と異なる．

Han ら [4] はジオタグ付きツイートにおける経緯度情報と単語情報をもとに，地理的な位置を示す単語 (Location indicative words; LIWs) を抽出する特徴選択手法を提案し，ツイートに含まれる単語情報から位置推定手法の精度向上を実現している．本研究が想定する状況においては，文書には正確な経緯度情報が付与されていないため，Han らの方法を直接用いることができない．例えばジオタグ付きツイート等の外部データとそこから抽出された LIWs を用いることにより，関連地名の位置情報推定精度を向上し，関連地名の抽出精度の向上が可能であると考えている．

3. 問題の定義

まず入力として，POI の集合 P と Web 文書の集合 D が与えられるものとする．POI p は，名称，緯度，経度をもつ．例えば，大さん橋は名称=“大さん橋”，緯度=35.45, 経度=139.65 をもつ．Web 文書 d は，テキストで表現される文書であり，例としてはブログ記事の本文がある．

出力は，各 POI に対する関連地名の集合である．例えば，「大さん橋」という POI について，「関内」という関連地名を得る．

関連地名とは，POI と関連のある場所の表現である．例えば，「関内」や「横浜中華街」のような表現は，「関内にある XX という店」のような形で POI の所在地を表すために用いることができ，XX という POI に対する関連地名ということができる．

4. 提案手法

本手法では，文書の集合 D に含まれる文書に対して固有表現抽出を行い，各 POI の名称と共起する地名の固有表現（LOCATION クラス）[5] を関連地名として抽出する．そして，POI と関連地名のペアについて，共起スコア，地理スコアの 2 種類のスコアを計算し，これらを基にしてその POI に対する関連地名の関連度を求める．最後に，各 POI について，関連度の高い関連地名を出力する．

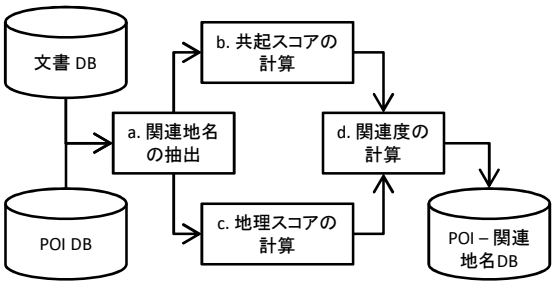


図 2 処理の流れ

図 2 は，本提案手法の流れを表したものである．ここからは，この図を参照して手法を説明する．

4.1 文書からの関連地名の抽出

ここでは，図 2 の a にあたる処理について説明する．

関連地名を抽出する対象の POI を POI の集合 P から 1 つ選び，p とする．まず，文書の集合 D について，p の名称を文字列として含む文書を検索し，それらの文書を D' とする．具体的には，データベース等に格納した文書に対して全文検索エンジンを用いる．

次に，D' に含まれる文書 d について，地名の固有表現を抽出する．まず，d に対して形態素解析と固有表現抽出を行う．IREX 日本語固有表現抽出タスクの定義では，固有表現には「人名（PERSON）」や「組織名（ORGANIZATION）」等 8 種類のクラスがあるが [5]，ここでは「地名（LOCATION）」を対象とする．

次に，d を文単位で走査し，p の名称を含む文の集合 S を得る．S に含まれる文 s について，地名の固有表現が含まれていれば，それを p に対する d における関連地名の集合 $N_{p,d}$ に追加する．

上記の操作を D' のすべての文書について行い，関連地名を集計すると，p と共起した関連地名とその共起回数が得られる．さらに，これらの処理を P のすべての POI について行うことで，各 POI に対する関連地名とその共起回数が得られる．

4.2 共起スコアの計算

共起スコアとは，ある関連地名が，ある POI に対してすべての文書中で同一文内で共起した回数をもとにしたスコアである．図 2 では b の処理にあたる．

$p \in P$ と共起した関連地名 g について，その共起回数を p と共起するすべての関連地名の共起回数の最大値で割ったものが共起スコア $s_c(p,g)$ となる．例えば，「大さん橋」という POI に対して「関内」という関連地名が全文書中で 50 回共起しており，最も多く共起した別の関連地名の共起回数が 100 回のとき，「大さん橋」に対する「関内」の共起スコアは 0.5 である．

上記の操作を P のすべての POI について行う．

4.3 地理スコアの計算

地理スコアとは，ある関連地名が，ある POI に対して地理的に関連しているかどうかを表すスコアである．図 2 では c の処理にあたる．4.1 項で抽出された関連地名には，実世界における位置が事前に与えられていないため，関連地名の位置を推定する必要がある．関連地名の位置の推定には，その関連地名と共起した POI の座標の平均を用いる．

4.1 項で得られた全ての関連地名の集合を G とし，ここから 1 つの関連地名を選び g とする．次に，g と共起したすべての POI を集め，P_g とする．

P_g のすべての POI について，それらの座標の平均を求め，これを g の位置 $c_g = (x_g, y_g)$ とする．そして，P_g の POI p について，次の値を求める．$dist(p, c_g)$ は p, c_g 間の距離を表す．ϵ は定数である．

$$s_g(p,g) = \frac{1}{dist(p,c_g) + \epsilon} \tag{1}$$

これを p に対する g の地理スコアとする．

4.4 関連度の計算

関連度とは，POI と関連地名が関連の強さを表すスコアであり，共起スコア，地理スコアを基にして求める．図 2 では d の処理にあたる．

ある POI p とある関連地名 g の関連度 $s_r(p,g)$ を次の式で求める．

$$s_r(p,g) = \lambda s_c + (1-\lambda) s_g$$

ここで，λ は係数であり，あらかじめ設定されているものとする．

5. 評価

提案手法の評価のために，2 つの実験を行う．実験 1 では λ の値による関連度の変化を観察し，実験 2 では人手で作成した正解データに基づく適合率を求める．

5.1 実験 1

POI の集合として神奈川県横浜市の POI を，Web 文書として日本語のブログ記事の本文を用いて関連地名の抽出を行い，λ の値による関連度の変化を見る．

5.1.1 実験条件

POI の集合として，神奈川県横浜市とその周辺に存在する POI 約 49,000 件を用いる．ジャンルとしては文化施設や小売，飲食店等が含まれる．Web 文書には，日本語のブログ記事を用いる．この中から各 POI について最大 200 件の記事を検索し，その本文を用いる．

上記のデータを用いて，各 POI について関連地名とその関連度を求め，関連度の高い順に上位 10 件の関連地名を求める．関連度の λ の値として $1.0, 0.75, 0.5, 0.25, 0.0$ を試行し，必要な場合はさらに中間の値を用いる．式 1 の地理

スコアは距離の単位を km とし，$\epsilon = 1$ とする．

5.2 結果と考察

特に POI が密集しており，日本語のブログにおいても多く言及されると思われる元町・中華街駅，関内駅，桜木町駅付近から選んだ 3 つの POI の結果を**表 1**, **表 2**, **表 3** に示す．括弧内は各 POI の所在地である．

元町・中華街駅付近から選んだ「福満園」(表 1) の $\lambda = 1.0$ の結果では，「横浜中華街」が 1 位に現れている．これは，中華街にある料理店としてブログ中で多く言及されているためであると考えられる．一方で，2 位以降には中国の地名が多く現れている．これらは固有表現抽出器が料理名等の一部を抽出したものと思われる．$\lambda = 0.5$ では地理スコアを考慮することにより，中国の地名の順位は下がっている．この例からわかるとおり，関連地名として抽出された地名の固有表現（LOCATION）の中には，関連地名として不適切なものが存在する．関連地名の抽出の再現率を担保するためには，ある程度の固有表現抽出誤りや，地名の固有表現を関連地名の候補として取得することによる偽陽性の発生は許容せざるを得ない．少なくともこの例では，地理スコアを用いて地理的な距離を考慮することによって，このような不適切な関連地名の一部のスコアを下げることが可能であることがわかる．

$\lambda = 0.0$ として地理スコアのみを考慮すると，横浜中華街も 11 位以下になり，別の店の名称（「菜香新館」）や非常に狭い範囲をもつ関連地名（「みなとみらい線日本大通り駅」）が現れる．これは，提案手法の地理スコアが関連地名が指す範囲の広さを考慮しておらず，「横浜中華街」のような広い範囲をもつ関連地名よりも，狭い範囲をもち，より POI に近いところにあるものに高いスコアを与えるためであると考えられる．

桜木町駅付近から選んだ「神奈川県立歴史博物館」(表 2) の $\lambda = 1.0$ の結果では，実際の所在地から離れた「鎌倉」が 1 位に現れている．$\lambda = 0.25$ でも 3 位であったため，さらに 0.1 まで下げたところ，表に示すように 11 位以下になった．$\lambda = 0.0$ では 0.1 とほとんど変わらないが，本来関連があるはずの「横浜」も 11 位以下になった．「鎌倉」と「横浜」の実際の共起回数はそれぞれ 25 回，24 回であり，その次に多い「鎌倉国宝館」の 15 回と比べてともに高く，共起スコアの影響が強く出ているといえる．「鎌倉」の共起回数がこれほど多い理由としては，博物館の展示内容への言及が多かったためであると考えられる．

関内駅付近から選んだ「ノーブル」(表 3) は，$\lambda = 1.0$ では東京都の関連地名が多数現れている．この傾向は $\lambda = 0.25$ でも変わらず，$\lambda = 0.0$ では少し減るものの全体として関連があるとは言い難い結果となった．これは，「ノーブル」という名称の POI が東京都を含む各地に存在するため，今回意図した関内駅付近の POI 以外に関する記述を多数検出したためであると考えられる．

表 3 POI「ノーブル」(関内駅付近) の関連地名

順位	$\lambda = 1.0$	$\lambda = 0.25$	$\lambda = 0.0$
1	有楽町	有楽町	フライブルク
2	横浜	横浜	ヒルズ
3	関内	関内	大船植物園
4	赤坂	フライブルク	横浜市中区吉田町
5	東京 23 区	ヒルズ	逗子
6	フライブルク	大船植物園	可愛
7	ヒルズ	赤坂	代々木上原
8	大船植物園	東京 23 区	上大岡
9	横浜市中区吉田町	横浜市中区吉田町	関内
10	逗子	逗子	米

5.3 実験 2

実験 1 の結果から，少数の POI については λ が関連度に影響を与えることがわかった．実験 2 では，より多くの POI について，人手で正解・不正解の判定を行った POI と関連地名のペアを用意し，上位 5 件における適合率を求める．

5.3.1 実験条件

POI の集合と Web 文書は実験 1 と同じものを用いる．正解データには，POI の集合から事前に 50 件の POI を取り出して関連地名の候補を求め，3 人の評価者が正誤を判定したものを用いる．これらの POI には，横浜中華街の飲食店，その他の飲食店，文化施設の 3 つのカテゴリから著名と思われるものを選んだ．各評価者は，POI p と関連地名 g のペアについて，p と g をキーワードとした Web 検索の結果の上位 10 件〜20 件を閲覧し，「p は g の POI である」，「p は g の近くの POI である」，「どちらにも当てはまらない」という判定を行う．

これらの判定結果から，2 人以上の評価者が「p は g の POI である」または「p は g の近くの POI である」と判定したペアを正解とした．例えば，「パシフィコ横浜」という POI について，関連地名「桜木町」と「関内」はそれぞれ 3 人，2 人が前述の判定をしたため正解とし，「元町」は 1 人のみであったため不正解とした．

上記のデータを用いて，実験 1 と同様に各 POI について関連度の高い順に関連地名を出力し，その上位 5 件を対象として適合率を計算する．ただし，正解データがないために正誤が判定できない関連地名は除外する．

5.4 結果と考察

λ の値ごとの適合率を**表 4** に示す．各カテゴリの POI の数は，中華街付近の飲食店が 7 件，その他の飲食店が 14 件，文化施設が 19 件であった．5 件以上の関連地名が得られなかった POI については計算の対象外とした．

全体の結果を見ると，$\lambda = 0.25$ が最も高い．これは，共起スコアと地理スコアの両方を考慮することが効果的であ

表 1 POI「福満園」（元町・中華街駅付近）の関連地名

順位	$\lambda = 1.0$	$\lambda = 0.5$	$\lambda = 0.0$
1	横浜中華街	横浜中華街	菜香新館
2	四川	菜香新館	SARIO 中華街
3	上海	SARIO 中華街	The CAFE 中華街
4	福建	The CAFE 中華街	萬珍樓売店中華街大通
5	湖南	萬珍樓売店中華街大通	楼
6	中華街	楼	重慶飯店新館
7	箱根	重慶飯店新館	招福門
8	みなとみらい線日本大通り駅	招福門	みなとみらい線日本大通り駅
9	重慶	みなとみらい線日本大通り駅	中華街駅
10	みなとみらい線元町	四川	赤レンガ

表 2 POI「神奈川県立歴史博物館」（桜木町駅付近）の関連地名

順位	$\lambda = 1.0$	$\lambda = 0.1$	$\lambda = 0.0$
1	鎌倉	みなとみらい線馬車道駅	みなとみらい線馬車道駅
2	横浜	馬車道駅	馬車道駅
3	鎌倉国宝館	馬車道通り	馬車道通り
4	金沢文庫	中華街	中華街
5	神奈川県立金沢文庫	横浜市中区	横浜市中区
6	馬車道	馬車道	桜木町駅
7	関内	桜木町駅	宝永
8	赤レンガ倉庫	宝永	JR 関内
9	かながわ	横浜	関内駅
10	桜木町	JR 関内	馬車道

表 4 カテゴリごとおよび全体の適合率

カテゴリ	$\lambda = 1.0$	$\lambda = 0.25$	$\lambda = 0.0$
中華街付近の飲食店	0.543	**0.686**	**0.686**
その他の飲食店	**0.800**	0.743	0.686
文化施設	0.811	**0.842**	0.779
全体	0.760	**0.780**	0.730

ることを示している．

次に，個々のカテゴリの結果を見る．中華街の飲食店では，$\lambda = 1.0$ のみが低い結果となった．これは，評価実験 1 の結果と同様に，料理等に関連して中国の地名が多く現れるためであると考えられる．一方，その他の飲食店では，$\lambda = 1.0$ が最も高く，共起スコアが重要であることを示している．文化施設では，$\lambda = 0.25, \lambda = 1.0$ の順に高くなっており，全体の結果と近い傾向を示している．

これらの結果から，総合的には共起スコアと地理スコアを両方考慮する $\lambda = 0.25$ 前後が良い結果を示す傾向があるといえる．また，中華街の飲食店とその他の飲食店の結果から，文書中での言及のされ方に関する POI の特性によって，共起スコアの重要性が異なると考えられる．

6. まとめ

本研究では，与えられた POI の集合に対して，POI の関連地名を Web 上の文書から抽出し，付与する手法を提案した．POI と関連地名の間に関連度を定義し，両者の同一文中での共起回数に加えて，関連地名の推定位置を用いた地理的な距離を同時に考慮した．評価実験では，関連度の高い順に上位 10 件の関連地名を出力して順位を比較し，さらに人手で作成した正解データを用いて上位 5 件の適合率を求めた．その結果，$0 \leq \lambda \leq 0.5$ で比較的良い結果が得られた．一方で，各地に同名のものが存在する POI や，他の地域とともに言及されやすい POI については抽出が難しいことがわかった．

今後の課題としては，関連地名の抽出と，地理スコアの計算の 2 点がある．関連地名の抽出においては，係り受け解析のようなより高度な処理を行うことによって，誤った固有表現を抽出することを防ぐことを検討する．地理スコアの計算については，関連地名が指す範囲の広さを考慮するように改良することで，範囲の広い著名な関連地名よりも，範囲が狭く，POI との距離が小さいものに高いスコアが与えられるという問題を解消できると考えられる．

参考文献

[1] 相良　毅, 喜連川優：Web からの効率的な新規店舗の発見・登録支援手法 (<特集>情報融合), 情報処理学会論文誌. データベース, Vol. 48, No. 11, pp. 49–57 (2007).

[2] 平野　徹, 松尾義博, 菊井玄一郎：地理的距離と有名度を用いた地名の曖昧性解消, 情報処理学会全国大会講演論文集, Vol. 70, No. 2, pp. 85–86 (2008).

[3] 手塚太郎, 近藤浩之, 田中克己：混合ガウス分布を用いたウェブコンテンツの地域性推定とオブジェクトレベルローカルサーチ, 情報処理学会論文誌. データベース, Vol. 1, No. 1, pp. 13–25（オンライン), 入手先 ⟨http://ci.nii.ac.jp/naid/110007989998/⟩ (2008).

[4] Han, B., Cook, P. and Baldwin, T.: Geolocation Prediction in Social Media Data by Finding Location Indicative Words, *International Conference on Computational Linguistics (COLING)*, Mumbai, India, p. 17 (2012).

[5] Sekine, S. and Isahara, H.: IREX: IR and IE evaluation project in Japanese, *Proceedings of International Conference on Language Resources & Evaluation* (2000).