D-019

# POI 情報を利用した Web 文書からの地名の抽出
# Extracting Geographic Names from Web Documents using POI Information

今井 良太† 廣嶋 伸章† 佐藤 隆† 鷲崎 誠司†
Ryota Imai Nobuaki Hiroshima Takashi Satou Seiji Susaki

## 1. はじめに

スマートフォンや高速モバイル通信の普及により，外出先で飲食店や観光スポットなどの Point of Interest (POI) を探す機会が増えている．例えば，近年のスマートフォンのほとんどに標準搭載されているマップ機能では，住所や現在位置によって指定した場所の POI を検索することができる．

POI の場所を示す表現には，場所との対応付けが明示的に行なわれ，一元的に管理されているものと，そうでないものが存在する．前者は，住所や郵便番号がそれにあたる．後者は，例えば周辺地域の人々の間でのみ通じる地名や，都市の開発によって新たに呼ばれるようになった地名を指す．本稿では後者の表現を「通称地名」と呼ぶ．

ある場所の表現に関係する POI を提示するシステムを考えたとき，管理されている表現については網羅的に処理することができる一方で，通称地名は場所との対応付けが明示的でなく，処理することが難しい．そのため，通称地名に対応する POI を事前に調べる必要がある．

本稿では，POI と地名の対応関係を調べる前段階として，通称地名を Web 上の文書から効率的に抽出することを目的とする．具体的には，実世界の特定の地点に存在することがわかっている POI の情報を利用し，文書中の固有表現の中からより地名らしい文字列を抽出する．

2 節では，本稿の関連研究について述べる．3 節では，本稿が扱う問題を定義する．4 節では，3 節の問題に対する提案手法について述べる．5 節では，提案手法に対する評価を行ない，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．

文書中の地名が指す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，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．

## 3. 問題定義

本稿で扱う問題では，Web 文書の集合と，それらの文書を絞り込むキーワードの集合を入力とし，通称地名と思われる文字列を出力とする．

Web 文書は，テキストで表現できるものを対象とする．キーワードは，Web 文書の集合を絞り込むためのものである．

## 4. 提案手法

本稿で提案する地名抽出手法は，以下の 3 段階からなる．

a) キーワードによる文書の絞込み
b) 文書に対する場所の固有表現抽出
c) 地名のフィルタリング

### 4.1 キーワードによる文書の絞込み

入力とする文書の集合 $D$ から通称地名を含む文書を絞り込むために，キーワードの集合 $K$ を用いる． $K$ に含まれるキーワード $k$ について，それを文字列として含む文書の集合 $D_k \subset D$ を得る．実際の絞込みには全文検索エンジンを用いる．

キーワードには，通称地名と共起する可能性の高いものを選定する．提案手法では，あらかじめ実在する POI の情報を収集しておき，これらの POI の名称をキーワードとして用いる． キーワードに実在する POI の名称を利用することで，他のキーワードを用いる場合と比べて，それと共起する地名も実在する可能性が高まると考えられる．

### 4.2 文書に対する場所の固有表現抽出

キーワード $k$ を含む文書 $d \in D_k$ について，形態素解析と固有表現抽出を行なう．ここから， $k$ と 1 つの文の中で共起する場所の固有表現を取り出し，これらを $d$ に対応する通称地名の候補 $G(k,d)$ とする．

上記操作を $k$ を含むすべての文書 $^{\forall}d \in D_k$ について行ない，

$$G(k,d_1), G(k,d_2), \ldots, G(k,d_m), \quad m = |D_k|$$

を得る．

さらに，ここまでの操作をすべてのキーワード $^{\forall}k \in K$ について行なう．最後に，すべての $G(k_n,d_m)$ $(n = |K|)$ に含まれる通称地名の候補の数を集計することで，すべてのキーワードによって得られた通称地名の候補とその出現回数が得られる．

### 4.3 地名のフィルタリング

4.2 の処理で得られた通称地名の候補には，住所の一部のような通称地名ではないものが含まれる．そのため，通称地名ではないことがわかっているものを除外するフィルタリングを行なう．具体的には，通称地名の各候補について，次のような条件にマッチするものを除外する処理を行なう．

- 都道府県（末尾が{都, 道, 府, 県}のいずれかである）
- 市区町村（末尾が{市, 区, 町, 村}のいずれかである）

† 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

- 番地（～丁目，数字とハイフンの組合せを含む）
- 国名
- 都市名
- POI の名称

## 5. 評価実験

POI の名称をキーワードとして用いる提案手法の有効性を検証するために，キーワードを POI のジャンル名に置き換えた場合との比較を行なった．

### 5.1 入力データ

次のような 2 つのデータセットを用意した．

a) POI 版（提案手法）

キーワード : POI の名称

下記ジャンル版で選定したジャンル名を用いて，「横浜市　<ジャンル名>」で Google の Web 検索を実行し，上位のページに掲載されている神奈川県横浜市の POI 160 件を手動で収集した．各ジャンルごとの POI の件数は表 1 のとおりである．

b) ジャンル版

キーワード : 横浜　AND　<ジャンル名>

Wikipedia の横浜市各区のページを参考に，4 種類のジャンルを選定した．ただし，飲食店については他のジャンルに比べて POI の数が多く，提供するメニューでジャンルが細分化されていることを考慮し，「横浜市　飲食店」で Google の Web 検索を実行し，上位のページを参考に 8 種類を選定した．キーワードとして用いる際は，ジャンル名のみでは横浜市以外の文書を含んでしまうため，「横浜」も同時に含むことを条件とする．具体的なジャンル名は表 1 のとおりである．

表 1: ジャンル名と POI の件数

| ジャンル | POI の件数 |
|---|---|
| 文化施設 | 20 |
| 公園 | 20 |
| 寺社 | 20 |
| レジャー | 20 |
| 居酒屋 | 10 |
| バー | 10 |
| カフェ | 10 |
| ファストフード | 10 |
| 創作料理 | 10 |
| 和食 | 10 |
| 中華 | 10 |
| 洋食 | 10 |

Web 文書には，日本語のブログ記事からキーワードごとに最大 100 件の記事を検索し，その本文を利用した．

### 5.2 評価方法

POI 版のデータを使用する提案手法と，ジャンル版のデータを使用する比較対象の 2 通りの出力結果を比較する．出力の総数は，POI 版は 3907 件，ジャンル版は 2697 件であった．得られた地名について，両方に出現する地名を除外し，残った地名から文書中での出現回数の多い順に上位 100 件ずつ，計 200 件を抽出した．これらの地名を，入力データを隠した形で 3 名の評価者に提示し，それが通称地名であるかどうかを判定してもらった．評価者には，横浜市に住んでいるか，最近まで住んでいた者を選定した．

### 5.3 結果と考察

表 2: 各評価者の判定による適合率

| データ | A | B | C |
|---|---|---|---|
| POI | 0.19 | 0.02 | 0.02 |
| ジャンル | 0.08 | 0.01 | 0.01 |

表 3: 地名の例

| データ | 地名 |
|---|---|
| POI | 中華街大通り，さくら通り |
| ジャンル | 横浜野毛 |

表 2 は，各入力データから得た地名のうち，評価者 A, B, C が通称地名と判定した数の割合を示したものである．各評価者において，提案手法である POI 版でわずかに高い適合率が得られた．評価者 A と他 2 名に差がみられるが，これは公園や商業施設等のうち，内部に別の POI をもつような規模の大きなものを地名ととらえるかどうかで差が出たものと考えられる．

表 3 は，実際に得られた通称地名の例であり，3 名の評価者全員が通称地名と判定したものである．POI 版の 2 件は通りの名前で，ジャンル版の「横浜野毛」は「横浜市中区野毛町」を指すものと思われるが，いずれも住所にはなっていない．

## 6. まとめ

本稿では，住所のように管理されていない場所の表現である通称地名を Web 上の文書から抽出する手法を提案した．この手法では，実在することがわかっている POI の名称を用いて文書を絞り込むことで，効率的に通称地名を抽出する．評価実験では，POI の名称の代わりに文化施設やファストフードといった POI のジャンル名を用いた場合の出力結果を比較対象として，人手によって評価した．

今後の課題として， POI と地名の対応関係や，POI 間の位置関係を利用することで，さらに抽出の精度を高めることが考えられる．その際には，チェーン店のような同名で複数の位置情報をもつ POI への対応が必要である．応用先としては，POI と地名の関係を地理情報のナビゲーションに活用することを検討している．合わせて，POI 情報の収集方法についても検討する必要がある．

### 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 (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 (2008).