D-023

昔の住所を持つ人物の地図上への表示

Displaying People with Old Addresses on a Map

張 鋼† 村上 晴美†
Gang Zhang Harumi Murakami

1. はじめに

個人の生活においては，先祖や親族を理解し，その情報を整理することは重要であり，彼らが過去や現在にいた（いる）場所を地図上に表示することは有用である．

本研究は，過去から現在にいたる人物にアクセスするための地図インタフェースの開発を目的とする．戸籍等の文書に含まれる情報から人物の過去及び現在の位置情報の取得を目指すが，記載された情報は，表記にゆれがある，地番と住居表示が混在している，現在は存在しない住所が記載されている，等の問題がある．本研究では，出生地等の文字列を入力として，Web 上の情報，郵便番号データ[1]，位置参照情報ダウンロードサービス[2]を利用して，現在の住所に変換する手法を提案する．得られた住所を Google Geocoding にかけて位置座標を取得する．

戸籍の出生年，没年，出生地を利用して，大正から現在にいたる親族を地図上に表示するシステムを試作した．年代の指定や歴史年表上のイベントからの選択により，指定した年代に生存している親族を地図上に表示できる．

以下，2 節で提案手法，3 節で評価実験，4 節で試作したプロトタイプ，5 節で関連研究について述べる．

2．提案手法

2.1 概要

提案手法は，情報収集部，判定部，補正部から構成される（図 1）．情報収集部では，対象住所の郵便番号候補を取得する．判定部は，郵便番号候補に対して，出現頻度，編集距離を利用してスコア付けを行い，最適な郵便番号を一件出力する．補正部は，位置参照情報ダウンロードサービスを利用して住所の精度を向上させる．

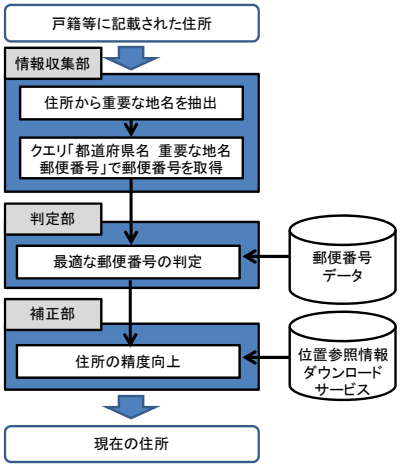

図 1 提案手法の概要

†大阪市立大学，Osaka City University

表 1 に提案手法の出力例を示す．本稿では，個人情報保護のため，番及び番地がある場合には 99 と記載している．以下では，No. 1 の例で説明する．

表 1 提案手法の出力例

No.	戸籍に記載された住所（手入力）	提案手法で出力された住所
1	大阪市南区内安堂寺町通 1 丁目 99 番地	大阪府大阪市中央区安堂寺町一丁目
2	大阪市北区北野小深町 99 番地	大阪府大阪市北区芝田一丁目
3	広島県芦田郡国府村字高木 99 番屋敷	広島県府中市高木町

2.2 Web 検索による郵便番号の取得

行政単位や地名は時代によって変遷する．本研究では，「重要な地名」と「現在の都道府県」に関連する郵便番号に着目し，候補として取得する．

2.2.1 重要な地名の抽出

住所の文字列を後ろからたどり，最後の数字の左を右端とし，2 字以降に最初に出現する「都，道，府，県，市，区，町，村，郡，または字」の右を左端とする文字列を抽出する．「大阪市南区内安堂寺町通 1 丁目 99 番地」は「内安堂寺町通」となる．

2.2.2 都道府県名の取得

クエリ「都道府県　番地等を除去した住所」で Web 検索を行い，上位 5 件のスニペットから，都道府県辞書との照合処理により，最頻度の都道府県名を抽出する．「都道府県　大阪市南区内安堂寺町通 1 丁目」というクエリになり，「大阪府」が出力される．

2.2.3 郵便番号の取得

クエリ「都道府県名　重要な地名　郵便番号」で Web 検索を行い，上位一件のスニペットから以下の正規表現を用いて郵便番号を抽出する．

```
\\d{3}-\\d{4}
```

上位一件で郵便番号が取得できなかった場合は，上位 5 件で同じ処理を行う．郵便番号が取得できなかった場合は提案手法は終了し「なし」と出力する．

例では「大阪府　内安堂寺町通　郵便番号」というクエリになり，上位一件から 542-0061，542-0067，541-0000 等の郵便番号が取得できる．

2.3 最適な郵便番号の判定

抽出した郵便番号を郵便番号データ[1]にかけて住所を取得する．郵便番号の出現頻度，郵便番号データに含まれる住所と，番地等を除去した元データの住所間の編集距離（レーベンシュタイン距離）を用いて，以下の式で郵便番号のスコア付けを行い，最上位の郵便番号を取得する．

$$score=\frac{f}{d} \quad (1)$$

ただし，f は抽出した郵便番号の出現頻度，d は郵便番号データから抽出した住所と，番地等を除去した元データの住所の編集距離を表す．

表 2 は判定例を示す．0.273 (3/11)が最上位スコアとなり，住所は「大阪府大阪市中央区安堂寺町」となる．

表 2　スコア計算例

郵便番号	郵便番号データ	t	d	$score$
542-0061	大阪府大阪市中央区安堂寺町	3	11	0.273
542-0067	大阪府大阪市中央区松屋町	1	16	0.063
541-0000	大阪府大阪市中央区	1	15	0.067

2.4 住所の精度向上

2.4.1 丁目の補完

郵便番号データに含まれる住所の文字列は，多くの場合丁目を含まないため，位置参照情報ダウンロードサービスを利用して情報の補正を行う．まず，元の住所に「丁目」がある場合は 2.3 で得た住所に丁目を結合する．ない場合は「一丁目」を結合する．次に，位置情報ダウンロードサービスにかけ，完全一致した場合にはその内容を住所とし，しなかった場合は「丁目」を除去する．例では「大阪府大阪市中央区安堂寺町一丁目」となる．

2.4.2 町名の除去

戸籍には「吹田市」「東京都渋谷区」のような記述がある．2.4.1 までの処理により不要な町名が結合された場合には町名以下を除去する．

2.5 位置情報の取得

提案手法で得た住所を Google Geocoding にかけて位置情報（緯度経度）を取得する．

3．評価実験

3.1 方法

親族データセット[3]の出生地と死亡地を対象とした．親族のため同じ文字列が多く，住所の異なり数は 32 件であった．番地等を除去した住所を役所で調べた住所を正解とした．東京都の区は市町村とした．比較手法として，番地等を除去した住所を Google Geocoding API v3 にかけて最上位の住所を取得する．性能は適合率と再現率で以下のとおり評価する．

$$適合率=\frac{r}{n} \quad (2)$$

$$再現率=\frac{r}{c} \quad (3)$$

ただし，r：出力した正解データ数，n：出力したデータ数，c：現在の住所の数とした．

3.2 結果

表 3 に実験結果を示す．特に再現率が向上しており提案手法の有効性を示している．

表 3　評価実験結果

	提案手法		比較手法	
	適合率	再現率	適合率	再現率
都道府県	100% (31/31)	97% (31/32)	96% (24/25)	75% (24/32)
市町村	87% (27/31)	84% (27/32)	80% (20/25)	63% (20/32)
町	64% (16/25)	64% (16/25)	60% (9/15)	32% (8/25)

4．プロトタイプ

戸籍の出生年，没年，出生地の情報を利用して親族を地図上に表示するプロトタイプを試作した．年代を入力するかスライダーを調整すると，該当年代に存在する人物を Google Map 上に表示する．実行例を図 2 に示す．

図 2　プロトタイプ

5．関連研究

FamilySearch[4]ではユーザが入力した都道府県レベルの情報で地図表示が可能であるが，より詳細な位置情報の推定は行われていない．Yamamoto ら[5] は Web 上の情報を用いて歴史的イベントの主な場所を推定して地図上に表示するが，主に都道府県や市町村レベルである．本研究では Web 上の情報と郵便番号に着目して昔の住所に合わせて都道府県，市町村，町レベルで変換する．

6．おわりに

住所の文字列を入力として，Web 上の情報，郵便番号データ，位置参照情報ダウンロードサービスを利用して，現在の住所に変換する手法を提案した．得られた住所を Google Geocoding にかけて位置座標を取得し，親族を地図上に表示するシステムを試作した．今後の課題として，住所変換の精度向上及び，出生地以外の所在地（たとえば本籍地）の利用や，地図上の人物関係の表示等があげられる．

参考文献

[1] http://www.post.japanpost.jp/zipcode/download.html
[2] http://nlftp.mlit.go.jp/isj/
[3] 鄭 寧, 村上 晴美, “家系図の視覚化：時系列の直系検索機能を持つ親族検索システム”, 2011 情報処理学会第 73 回全国大会 (2011).
[4] FamilySearch, http://www.familysearch.org/
[5] Yamamoto, M., Takahashi, Y., Iwasaki, H., Oyama, S., Ohshima, H., and Tanaka, K., “Extraction and Geographical Navigation of Important Historical Events in the Web”, SAC '13 Proceedings of the 28th Annual ACM Symposium on Applied Computing, pp. 878-885 (2013).