Webページからの人物に関する位置情報の抽出

高守 雄也† 上田 洋‡ 村上 晴美§

†§大阪市立大学大学院創造都市研究科 ‡大阪市立大学大学院工学研究科

1. はじめに

近年のblogやSNSの普及によりWeb上で情報発信する人々が飛躍的に増加している．それに従い，Web上に登場する同姓同名人物の数も多くなってきている．Web上の人名検索において同姓同名人物を識別する問題は重要となってきている．このような背景から，Web上の同姓同名人物の識別に関する研究が盛んに行われている．それらの多くは，[1]のように人名検索結果であるWebページを人物毎にクラスタリングする研究である．しかし，ただクラスタに分類するだけでは各クラスタが誰であるのか認識するためには，人物毎に分類されたWebページずつ閲覧しなければならず，ユーザにとって負荷が高い．ユーザが求める人物を簡単に選択するためには，ただWebページを分離するだけではなく，同姓同名人物を識別するための簡単なインタフェースが必要である．

本研究では，氏名によるWeb検索の結果，同姓同名人物毎に分けられたWebページ群に対して該当人物を表すオブジェクトを地図上に表示することを目的とする．そのために，該当人物に適切な位置情報（本研究では位置情報ラベルと位置座標の組とする）を一つ付与することを目的とする．

本稿の構成は以下のとおりである．2節で提案手法について述べる．3節では提案した手法の評価実験について述べる．4節では関連研究と比較する．

2. 提案手法

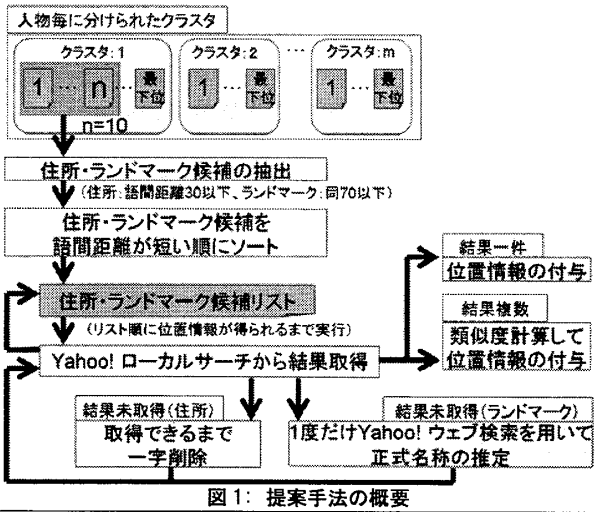

図1：提案手法の概要

Extracting location information about people from Web pages
†Yuya Takamori §Harumi Murakami
Graduate School for Creative Cities, Osaka City University
‡Hiroshi Ueda
Graduate School of Engineering, Osaka City University

本研究では，該当人物に最適な位置情報ラベルと位置座標を付与するために，Webページに含まれる住所やランドマークを抽出して，Yahoo!ローカルサーチAPI（以下，Yahoo!ローカルサーチ）を用いる．主要なアイディアは，ランドマークへの着目，語間距離と検索順位の利用である．

提案手法は，(1) 住所・ランドマーク候補リストの作成，(2) 位置情報の取得，に大別される．図1に提案手法の概要を示す．

2.1 住所・ランドマーク候補リストの作成

2.1.1 住所・ランドマーク候補の抽出

氏名による検索結果の上位10件のWebページを対象としてMecabにより形態素解析を行い，以下の処理にて抽出する．

(1) 住所候補抽出

(a)，(b)，(c)の条件が2回以上連続している場合に接続して抽出する．

(a) 品詞名に地域，括弧開，括弧閉，形容動詞語幹を含む
(b) 候補語自体が日本語であり「地域」ではない
(c) 品詞名に地域が無い場合，候補語が「高専」「大学校」「通り」「大字」「小字」「東」「西」「南」「北」「-」であるか，品詞名に接尾または数がある

(2) ランドマーク候補抽出

(a)，(b)，(c)の条件がすべて満たされている場合に抽出する．

(a) 品詞名に組織が含まれている
(b) 品詞名に地域，人名が 含まれていない
(c) 候補語が日本語かつ1文字以上である
(d) 候補語自体が「組織」と言う名前ではない

2.1.2 氏名と住所・ランドマーク候補の語間距離計算

まず，住所・ランドマーク候補から，「日本」「本州」「九州」「四国」などの単位の大きすぎるものを取り除く．次に，Webページ中の氏名と住所・ランドマーク候補の語間距離を文字数にて計算する．住所候補は30以下，ランドマーク候補は70以下のものを取得する．最後に，語間距離が短いもの順にソートして住所・ランドマーク候補リストを作成する．

ソートの際，本文に含まれる候補から処理し，その後にタイトルに含まれる候補を追加する．

2.2 位置情報の取得

住所・ランドマーク候補リストの順番に，Yahoo!ローカルサーチにかけ，パターンマッチに基づき位置情報ラベルと位置座標を取得する．結果が得られた場合には解として処理を終了する．

結果を1件取得した場合には解となる.

結果が複数あった場合，住所・ランドマーク候補と住所・ランドマーク候補が抽出されたWebページを tf-idf 法で重み付けを行い，ベクトル空間モデルの余弦を用いた類似度計算を行う．その値が 0.7 以上で最も類似する結果を解とする.

パターンマッチにより結果が得られなかった場合，住所候補の場合は一字削除，ランドマーク候補の場合は正式名称推定を用いて，Yahoo!ローカルサーチとのパターンマッチを試みる.

2.2.1 一字削除

住所候補を，パターンマッチに成功するまで語尾の1字を取り除き，Yahoo!ローカルサーチにかける．取得できない場合は終了する.

2.2.2 正式名称推定

Yahoo!ローカルサーチでは，ランドマークは正式名称でなくてはならない．そこで，ランドマーク候補に正式名称がある場合には推定して変換する．ランドマーク候補を Yahoo!ウェブ検索 API（以下，Yahoo!ウェブ検索）にかけ，最上位のタイトルを正式名称とする．この処理は1回だけ実行し，結果が取得できない場合は終了する.

3. 評価実験

評価用データセットを作成した．まず，Yahoo!ウェブ検索より 20 の氏名[1]をクエリとして上位 100 件のデータを取得した．次に，取得したデータ(HTML ファイル)を人手で人物毎に分類し，Web ページから正解を抽出した．原則として，該当人物が現在いる場所（自宅または勤務先）を正解とした．正解が複数ある場合(例:「大阪市立大学」「大阪市住吉区杉本 3-3-138」「大阪市住吉区」にはどれを選択しても良いとした.

Web ページ群から住所・ランドマークを抽出する範囲を1件，5件，10件，15件，全件に設定し，精度，再現率，F値で評価を行った.

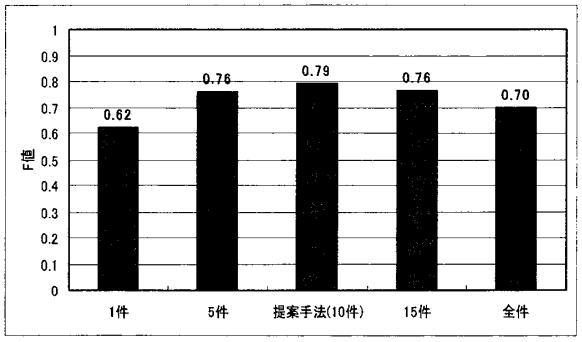

図2: Webページの件数

Webページの上位10件までを利用する提案手法は，精度 0.79，再現率 0.78，F値 0.79 であり，他の件数と比べて最も良かった（図2）．Web検索エンジンの検索結果，上位10件は住所・ランドマーク候補を抽出するページ数として適当であると考える.

また，ランドマークは抽出せず住所だけを抽出する比較手法（「住所」と呼ぶ）と，語間距離ではなく頻度を用いて候補を選択する比較手法（「頻度」と呼ぶ）との比較実験を行った.

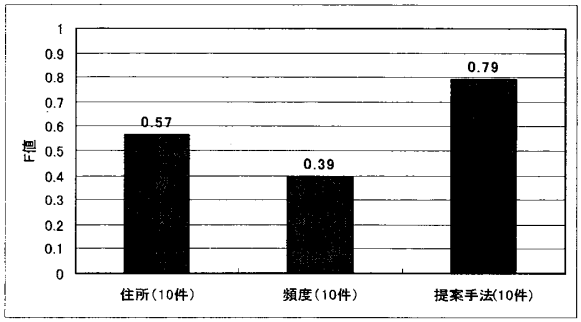

図3: 他手法との比較

住所手法はF値 0.57，頻度手法はF値 0.39 であり（図 3），提案手法が最も良かった．ランドマークと語間距離に着目した提案手法の有効性を示していると考える.

4. 関連研究

同姓同名人物を識別するために，人物毎に分類されたクラスタに適切なラベルを付与する研究が行われている．代表的な先行研究として Wan ら[2]は人名検索を行い，人物毎に Web ページを分類してクラスタを作成し，Web ページから抽出した語を用いてクラスタに肩書のラベル付けを行っている．上田ら[3]は分離された同姓同名人物に対して職業関連情報のラベルを付与している．本研究はクラスタに住所ラベルを付与する事に相当する.

Web ページ内に含まれる住所などから位置座標に変換することにより地図上にオブジェクト表示する研究が増えてきている（たとえば[4]）．本研究はランドマークに着目している点が[5] に関連する.

5. おわりに

Web 人名検索の結果，同姓同名人物に分けられた Web ページ群に対して，最適な位置情報ラベルと位置情報を付与する手法を提案し，インタフェースを試作した．実験の結果はF値 79%と良好であった.

参考文献

[1] 佐藤進也，風間一洋，福田健介，村上健一郎: 実世界指向 Web マイニングによる同姓同名人物の分離，情報処理学会論文誌: データベース，Vol. 46, No. SIG 8 (TOD26), pp. 26-36 (2005).

[2] Wan, X., Gao, J., Li, M. and Ding, B.: Person Resolution in Person Search Results: WebHawk, *Proceedings of CIKM2005*, pp. 163-170 (2005).

[3] 上田洋，村上晴美，辰巳昭治: Web 上の同姓同名人物識別のための職業関連情報の抽出，人工知能学会全国大会(第 22 回)論文集 (2008).

[4] 新井イスマイル，川口誠敬，藤川和利，砂原秀樹: 個人サイトの評価情報と位置情報に基づいた店舗検索用 Web インデサの開発，情報処理学会論文誌，Vol. 48, No. 7, pp. 2319-2327 (2007).

[5] 細川宜秀，高橋直久: ドキュメント・データを対象としたジオ・コーディング手法，情報処理学会研究報告，No. 2003-DBS-130, pp. 87-93 (2003).