3E4-3in

Web 上の人物履歴情報の地図表示システム

Map Display System for Curriculum Vitae on the Web

唐 春亮[*1] Chunliang Tang　王 爽[*1] Shuang Wang　上田 洋[*2] Hiroshi Ueda　村上 晴美[*1] Harumi Murakami

[*1] 大阪市立大学大学院創造都市研究科 Graduate School for Creative Cities, Osaka City University

[*2] 株式会社 ATR Creative ATR Creative

We present a system that displays curriculum vitae on a map to understand people. Our method is based on the following processes: (1) creating curriculum vitae using related work [Ueda 10], (2) extracting place names where the person studied and worked from the vitae, (3) getting latitudes, longitudes and addresses from the place names using Google Maps API, and (4) displaying a vitae along with a map using Google Maps.

1. はじめに

Web 人物検索の重要性が増大しているが，同姓同名人物を含む大量の人物を判別，理解するためには人物に関連する多様な情報が必要である．

本研究では，Web 上の人物検索における地図インタフェースの開発を目的とする．Web 上から人物の過去，現在の所在地の抽出を目指すが，所在地として人物が一定期間いた場所が重要であると考え，人物の履歴情報の中で学歴や職歴に着目して学校や勤務先の所在地を取得する．先行研究[上田 10]において，Web 検索エンジンで氏名を入力して履歴書を作成する手法を開発した．本研究ではこの手法を用いて履歴書を作成し，そこから学校と勤務先を抽出する．

以下，2 節では提案手法，3 節では試作したプロトタイプ，4 節では評価実験について述べる．

2. 提案手法

2.1 概要

提案手法の概要は以下のとおりである．2.2 で[上田 10]の手法を用いて履歴書を作成し，2.3 で履歴文（時間と，人物に関する出来事の両方を含む文）から学校と勤務先を抽出し，2.4 で抽出した学校と勤務先を Google Maps API にかけて位置情報を取得して位置情報を付与した学歴と職歴データを作成し，2.5 で地図表示を行う．

2.2 履歴書の作成

[上田 10]では，Web 人名検索結果の Web ページから履歴文を抽出し，4 つのカテゴリ（戸籍，学歴，経歴，受賞歴）毎に分類し履歴書の形式で提示する．経歴は職業に関連するものが多いが，出演した作品や，本業以外の活動なども含んでいる．手法は，(1) Web からの人物に関する履歴文の抽出，(2) 履歴文の分類，(3) 同義の履歴文のクラスタリング，の 3 つの処理から構成される．

(1) では，ヒューリスティックを用いて Web ページから履歴文を取得する．その後，不要語と不要パターンを用いてフィルタリングを行い，主に HTML のタグの出現パターンを用いて検索対象の人物と関係のある履歴文かどうか SVM を使用して判定する．(2)では，人物と関係があると判定された履歴文について，形態素解析を行い形態素と形態素数をパターンとし，SVM を用いて 4 つのカテゴリに分類する．(3)では，記述された時間が同じで内容も同義である履歴文をまとめる処理を行う．

本研究では，[上田 10]で作成される履歴書の中，学歴カテゴリから学歴，経歴カテゴリから職歴を作成する．

2.3 学校と勤務先の抽出

形態素解析とヒューリスティックを用いて，学歴と経歴カテゴリの履歴文から学校と勤務先を抽出する．

2.3.1 学校の抽出

1. 履歴文を学歴のカテゴリから指定したキーワード（「学」「校」または「卒」）を含む履歴文を抽出する．
2. MeCab で履歴文を形態素解析する．
3. 履歴文に「名詞」の「固有名詞」の「組織」または「地域」が存在する場合
   - 「組織」または「地域」を始点として名詞を連結する．
   - 学校の種類に応じて学校を抽出：大学院の（「大学院」を含む）場合は「研究科」を終点，大学の（「大学院」を含まず「大学」を含む）場合は「学部」を終点，それ以外は「学校」または「学院」を終点とする．（例：宇部市立琴芝小学校）
4. 「名詞」の「固有名詞」の「組織」また「地域」が存在しない場合には，「研究科」「学部」「学校」または「学院」を始点としてさかのぼって名詞を連結する．（例：麻布高等学校）
5. 同義の学校のフィルタリング：「入学」「編入」「卒業」（「卒」を含む）毎に分類し，抽出した学校を比較して，すべての文字が他の学校に含まれる学校を削除する．（例：東京都立小山台高等学校と都立小山台高校を比較して後者を削除する）

2.3.2 勤務先の抽出

1. 履歴文を経歴のカテゴリから指定したキーワード（政治家：「当選」「就任」「辞任」；研究者：「大学」「学校」「研究」；スポーツ選手：「入団」「移籍」「引退」；芸能人：「務め」「担当」；一般人：「就職」「入社」「退社」）を含む履歴文を抽出する．
2. MeCab で履歴文を形態素解析する．
3. 職歴を含まない履歴文のフィルタリング：履歴文に該当人物とは異なる「固有名詞」の「人名」の「姓」が存在して，

連絡先：唐 春亮，大阪市立大学大学院創造都市研究科，m12ucN5Q15@ex.media.osaka-cu.ac.jp

その後に「は」または「が」が存在する場合，その履歴文を除外する．(例:「中尾則幸は 1998 年第 18 回参議院銀に通常選挙で当初は(後略)」を削除する)

4. 「名詞」の「固有名詞」の「組織」が存在する場合
   - 「組織」を抽出する．(例:衆議院)
   - 複数の組織がある場合には，指定したキーワードに最も近いものを選択する．
5. 「名詞」の「固有名詞」の「組織」が存在しない場合
   - 「(株)」や「株式会社」が含まれている場合には，「(株)」や「株式会社」前後の「名詞」を連結する．(例:株式会社三笠興産，呉羽紡績株式会社)
   - 「大臣」が含まれている場合には，大臣辞書を用いて大臣名を抽出する．(例:運輸大臣)
   - 「市長」や「知事」が含まれている場合には，最も近い「名詞」の「固有名詞」の「地域」を抽出して，市長の場合は「市」，知事の場合は「都道府県」を付ける．(例:阿久根市)
6. 「大学」「学校」「研究」を含む場合(研究者とみなされる場合)には，2.3.1 学校の抽出の 3, 4 と同じ(ただし 4 は「研究所」「センター」からさかのぼる)．(例:奈良先端科学技術大学院情報科学研究科，国立情報学研究所)

2.4 位置情報の取得

抽出した学校と勤務先を Google Maps API にかけて最上位の位置情報(緯度経度)とラベルと住所を取得する．得られた位置情報とラベルと住所，元の学歴と経歴カテゴリの中の学校や勤務先を含む履歴文をファイルに書き込み，学歴と職歴データを作成する．

2.5 地図上に学歴と職歴の表示

学歴と職歴データを Google Map にかけて，マーカーを地図上に表示する．

3. プロトタイプ

氏名を入力して人物履歴情報を地図上に表示するプロトタイプを試作した．「菅　直人」氏での実行例を図 1 に示す．[上田 10]で作成された履歴書から，学歴はほぼ同じ履歴文を抽出し，職歴は経歴の中から選択的に履歴文を抽出している．

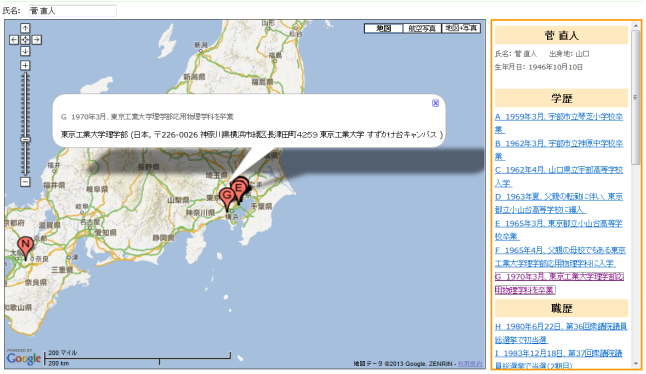

図 1:プロトタイプ

4. 評価実験

4.1 データセット

有名人を中心とする 56 人物(政治家:17 人，スポーツ選手 14 人，芸能人 12 人,研究者 10 人，企業家人 1 人，漫画家 1 人，歴史上の人物 1 人)の氏名を入力して Web 検索エンジンから人物毎に 50 件の HTML ファイルを取得した．[上田 10]の手法を実装して作成した履歴書中から抽出すべき学校と勤務先を人手で判定して正解データとした．

4.2 学校と勤務先の評価

学校と勤務先の抽出性能の評価を以下のとおり行った．

$$\text{適合率}=\frac{r}{n} \qquad \text{再現率}=\frac{r}{c}$$

ただし，r:抽出した正解データ数，n:抽出したデータ数，c:正解データ数とした．

表 1 に結果を示す．p_1, r_1 は完全正解のみを正解とする場合，p_2, r_2 は部分一致を正解に含む場合である．

表 1　学校と勤務先の抽出の結果

	適合率		再現率	
	p_1	p_2	r_1	r_2
学校	0.91 (59/65)	0.97 (63/65)	0.84 (59/70)	0.90 (63/70)
勤務	0.65 (154/241)	0.73 (166/241)	0.70 (154/233)	0.78 (166/233)

4.3 学歴と職歴の位置情報の評価

抽出した学校や勤務先の中，正解データを Google Maps API にかけて最上位が正しいかどうか評価したところ，学歴が 0.90(53/59)，職歴が 0.53(81/154)であった．

5. 関連研究

Web 人物検索結果から属性情報を抽出するコンテスト WePS-2 では，学校は評価対象に含まれたが，勤務先は曖昧であるとして省かれた[Sekine 09]．[Kimura 07]では Web 上の人物の経歴情報を抽出して時系列に提示している．本研究では，Web 上の人物の学歴と職歴として学校と勤務先を抽出して地図上に表示している．

6. おわりに

Web 上の人物の学歴や職歴を抽出し，地図上に表示する手法を提案した．[上田 10]で履歴書を作成した後に，ヒューリスティックを用いて学校と勤務先を抽出して地図上に表示する．今後の課題として，職歴(勤務先)の抽出性能向上，学校や勤務先からの位置情報獲得の性能向上などがあげられる．

参考文献

[上田 10] 上田 洋，村上 晴美，辰巳 昭治:Web 上の人物理解のための履歴書作成，人工知能学会論文誌，Vol.25, No.1, pp. 144-156, 2010.

[Sekine 09] Sekine, S., Artiles, J.: WePS2 Attribute Extraction Task, *2nd Web People Search Evaluation Workshop (WePS 2009), WWW 2009* , 2009.

[Kimura 07] Kimura R., Oyama, S., Toda, H., Tanaka, K.: Creating Personal Histories from the Web using Namesake Disambiguation and Event Extraction, *ICWE 2007*, pp. 400-414, 2007.