様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 ２１ 日現在

機関番号：２４４０２
研究種目：基盤研究(C)（一般）
研究期間：2013～2015
課題番号：２５３３０３８５

研究課題名（和文）Web上の人物を要約する手法の開発

研究課題名（英文）Development of Methodology for Summarizing People on the Web

研究代表者
村上　晴美（Murakami, Harumi）

大阪市立大学・大学院創造都市研究科・教授

研究者番号：４０３０５６４４

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：研究の目的はWeb上の人物を選択するための方法の開発である。主要な成果は以下の2点である。(1) Web上の人物にNDCの分類記号を付与する手法を提案した。関連する語（索引語）と分類記号を関連付けるNDCの相関索引を利用し、Webページのタイトル要素のテキストに含まれる索引語の頻度に基づき分類記号を求める。4つの手法と6つの文書を組み合わせた評価実験により、提案手法の有効性を確認した．(2) Web上の人物の履歴書と地図を表示するシステムを提案し、地名を抽出するアルゴリズムと位置情報に変換する手法の評価を行った。

研究成果の概要（英文）：The aim of this research is to develop methodology that helps users select people on the Web. The following are its main results. (1) We present a method of assigning Nippon Decimal Classification (NDC) to people on the Web. We use a relative index in NDC, which lists the related index terms attached to NDC. We count the number of relative index terms contained in the titles of Web pages. We evaluated the usefulness of our approach by comparing four methods and six documents and found that our method outperformed other methods and documents. (2) We presented a system that displays a curriculum vitae with a map for people on the Web, and evaluated our algorithms that extract place names and convert them into location information from Web search results.

研究分野：情報検索

キーワード：情報検索　インタフェース　NDCディレクトリ　履歴書　地図

１．研究開始当初の背景

Web 上の人物検索においては、人名の曖昧性解消が重要な課題となってきている。関連研究の多くが曖昧性解消（人物毎に Web ページの自動分類）を目指すが、本研究の目的は分類された人物の選択の支援である。

Web 上の人名の曖昧性解消の最近の動向は、(a) 曖昧性解消技術の精度の向上と、(b) 人物属性情報の抽出に大別されるが、本研究は(b)に関連する。情報抽出は、あらかじめ設定したターゲットである情報をすべて抜き出す技術であるが、人物選択のインタフェースに応用する場合、抽出したすべての情報を表示すると煩雑になり使いにくい。そこで、本研究では「人物の選択に有用な情報を抽出、生成、または付与し（本研究では「要約」と呼ぶ)」、インタフェースを開発する。

本研究は科学研究費（No.22500219）の助成を受けた先行研究を発展させるものである。先行研究の主な成果は「Web 上の同姓同名人物の分離過程の解明」と「NDC を用いた人物ディレクトリの開発」であった。

２．研究の目的

研究の全体構想は「Web 上の人物ディレクトリの開発」であり、「Web 上の人物を選択するためのインタフェースの開発」を目的とする。

具体的には人物の「要約」手法と 3 種類のインタフェース（表、NDC ディレクトリ、履歴書＋地図）を開発する。本研究における要約とは人物を選択・理解するために有用な情報の抽出、生成、あるいは付与である。

以下では、主要な研究成果として、(1) NDC の付与手法と NDC ディレクトリ、(2) 履歴書＋地図インタフェースについて述べる。

３．研究の方法

(1) NDC の付与手法と NDC ディレクトリ

先行研究で開発した NDC 付与手法と NDC ディレクトリの評価実験を行った。先行研究でも評価実験を行ったが、より詳細に分析をするために追加実験を行った。実験結果の統計分析も行った。

20 の日本人の氏名を用いて Web 検索を行い、20 氏名×100HTML ファイルの結果を取得し、人手で同姓同名人物に分離したデータセットを利用する。著名人や専門性のある人物が存在するが、専門性のない人物や架空の人物の存在も含め、152 人物が存在する。

先行研究の評価実験も含めて 5 つの実験を行い、関連研究の整理と議論を行い、学術論文にまとめた。

(2) 履歴書＋地図インタフェース

56 の日本人の氏名を用いて（ターゲットである人物の同姓同名人物を分離するために必要に応じて AND 検索としながら）Web 検索を行い、56 氏名×50HTML ファイルの結果を取得して作成したデータセットを利用する。著名人を含む専門性のある 56 人物が存在する。

先行研究[上田 2010]（上田、村上、辰巳：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、2010）の手法を用いて人物の履歴書を作成し、そこから人物の学歴と職歴を抽出してその所在地を地図上に表示する。

提案手法の有効性を実験により評価した。

４．研究成果

(1) NDC の付与手法と NDC ディレクトリ

①評価実験

以下では主要な実験である実験 1 について述べる。

NDC を人物（人物クラスタ）に付与する提案手法の有効性を、人手で付与した正解データに対する適合率、再現率、F 値、総合精度の 4 つの指標で検討した。

4 つの手法と 6 つの文書を用いて人物に NDC を付与した。4 つの手法とは、情報検索の一般的なベースラインである(a)Tf、(b)Tf-idf、(c) 余弦、および、(d)提案手法であり、6 つの文書とは、(1)タイトル、(2)全文、(3)スニペット、(4)Kwic50、(5)Kwic100、(6)Kwic200 である。ただし Kwic は氏名の前後の文字列であり、数字は前後の文字数を表す。

比較手法である(a)Tf、(b)Tf-idf、(c)余弦では、相関索引を使用せずに、分類項目名と文書に形態素解析 MeCab（標準の IPA 辞書を使用）をかけて 2 字以上の名詞を抽出し、提案手法と同じ不要語を除去したものを語として直接照合することとした。

完全一致について、提案手法においては、適合率ではタイトル、再現率ではスニペットか Kwic、F 値では Kwic50、総合精度ではタイトルが最も良かった。統計分析を行った。データが正規分布ではないため、手法と文書の要因ごとにノンパラメトリックな方法として、生データが完全に対応するものはフリードマン検定、対応しないものはクラスカル・ウォリス検定を行った。

手法については適合率、再現率、総合精度において有意差が見られた。適合率についてはクラスカル・ウォリス検定で有意差が見られ（$\chi 2(3)=42.681, p<.01$）、シェッフェ法で提案手法と他のすべての手法との間に有意差が見られた（$p<.01$）。再現率についてはクラスカル・ウォリス検定で有意差が見られ（$\chi 2(3)=29.814, p<.01$）、シェッフェ法で提案手法と他のすべての手法との間に有意差が見られた（Tf と余弦とは $p<.05$、Tf-idf とは $p<.01$）。総合精度についてはフリードマン検定で有意差が見られ（$\chi 2(3)=18.437, p<.01$）、シェッフェ法で提案手法と他のすべての手法との間（$p<.01$）、Tf-idf と余弦との間（$p<.05$）に有意差が見られた。

文書については、総合精度についてフリー

ドマン検定で有意差が見られ（$\chi 2(5)=14.597, p<.05$）、シェッフェ法で、タイトルと Kwic50 以外の文書との間（$p<.01$）、全文と Kwic との間（Kwic50 と 100 とは $p<.01$,Kwic200 とは $p<.05$）に有意差が見られたが、適合率と再現率については有意差は見られなかった。

以上より、提案手法の比較手法への優位性および、総合精度におけるタイトルの他文書への優位性を確認した。基本的に、タイトルの適合率の良さはゴミの少なさ、タイトル以外の再現率の良さはテキスト量の多さによると考える。適合率、F 値、総合精度においてタイトルの数値が良いことから、タイトルが第一選択肢と考える。

ただし、正解なしの 33 人をデータセットから除外して 119 人について計算すると、総合精度は再現率と同じとなり適合率以外の優位性は低下する。正解のある 119 人について統計分析を行った。手法については、適合率でクラスカル・ウォリス検定で有意差が見られ（$\chi 2(3)=38.015, p<.01$）、シェッフェ法で提案手法と他のすべての手法との間に有意差が見られた（$p<.01$）が、再現率および総合精度では有意差が見られなかった。文書については適合率、再現率、総合精度において有意差が見られなかった。

提案手法でタイトルを使う場合、完全一致において適合率が 0.18、再現率が 0.13、F 値が 0.15、総合精度が 0.28 と一見あまり良くない数値であるが、ラベルとして用いる場合には最適解でなくても人物に部分的に適合すれば多くの場合役に立つこと、スコア最上位ではない解が最適解である場合があること、実際にディレクトリで閲覧する際には第二次、第三次レベルで適合すればよいこと、などにより実用可能性がある。

②議論

評価実験の結果、Web 上の人名検索結果の HTML ファイル上位 100 件に含まれる人物に NDC を付与して人物ディレクトリを開発するために、索引語との照合に基づく手法を用いた場合、6 つの文書の中ではタイトルが最も良いことがわかった。

本研究の最大の貢献は、図書館の分類記号を Web 上の人物に付与して人物ディレクトリを開発したことにある。われわれの知るかぎり、本研究は Web 上の人物に図書館の分類記号を与える最初の研究である。

また、人物を表現するための疑似的な文書として、全文や Kwic 文書やスニペットよりもタイトルが良いことを示したことも貢献である。Kwic 文書は一般に情報検索の分野でウインドウサイズとよばれる概念（指定した文字列の前後の単語数）とほぼ同じである。日本語は何を単語とするかについて曖昧性があるため本研究では文字数を使用した。人物検索の中でも専門家を探すという典型的なタスクであるエキスパート検索では文書中の用語を抽出するためにウインドウサイズを用いることが多い。本研究で Kwic 文書よりもタイトルが良いという結果はエキスパート検索と異なる。この結果はエキスパート検索と Web 上の人物検索というタスクの違いを表しており、人物検索に新しい示唆を与えると考える。エキスパート検索は専門家という一定水準以上の情報量がある人物群の中で優れた人物を探すタスクであり、多くの場合対象となる情報源も均質である。本研究のタスクは Web という不均質な情報源と、多くが無名人で匿名の人物も含む Web 上の人物を対象としている。また、本研究のアプローチはテキストからの自然語の抽出ではなく、図書館の用語という統制語の付与である。さらに、本研究はユーザが使用するディレクトリ開発を目的としている。これらの背景より、ゴミとなる NDC を付与しないことが重要であり総合的にタイトルが良かったと考える。ただし、ページ数の多い人物を対象とする場合には実験 1 において Kwic 文書の結果が良かった。ページ数が多い人物の場合はエキスパート検索と類似した結果になる可能性がある。

本研究の適用可能性について述べる。本研究は NDC と日本語で行っているが、提案手法は DDC のように相関索引を持ち構造の類似した分類システムにはほぼそのまま適用可能である。相関索引を持たない分類システムの場合は、比較手法や関連研究で提案されている手法を実装すればよい。関連研究には類似文書に付与された件名標目や分類記号を参考にして機械学習のアプローチにより分類記号を付与するものがあるが、本研究のアプローチでは参考となるデータがない状況で新規に分類記号を付与することができる。本研究のアプローチは、自動分離したクラスタに対しても適用できる。また、同姓同名人物の分離だけではなく、Web から作成した人物データベース検索にも適用可能であると考える。さらには、Web 以外の人物に関連するテキストにも適用可能であると考える。ただしいずれの場合も、タスクに応じた文書と手法の選択・調整はするべきである。

今後のおもな課題は以下のとおりである。まず、提案手法には改善の余地がある。多義の索引語がある場合（索引語から複数の NDC に変換できる場合）、本研究ではすべてに変換したが、前後の文字列から文脈を判定するなどして重みを変えることが考えられる。また、本研究では相関索引以外の語の拡張を行わずにどこまでできるかを調べたが、今後は NDC の階層構造をはじめとする語の拡張も検討する必要がある。利用する情報を変えた場合には単純なパターンマッチよりも比較手法（余弦など）の方が良い可能性がある。また、現在の提案手法のままで、ページ数の多い人物と少ない人物で用いる文書を変えることが考えられる。つぎに、異なる種類のデータセット（例：有名人など）で評価実験

を行うことが考えられる。Web 検索エンジンの検索結果は時代とともに変わるため、経年変化を観察する必要もある。さらに、本研究は Web 人名検索結果を同姓同名人物に自動分離すること自体は研究の対象としていないが、自動分離したクラスタでどの程度使えるか検討することが考えられる。

最後に、本研究の新規性と有効性を整理する。まず、Web 上の人物検索において、同姓同名人物の分離後の人物の選択という問題に着目している。Web 上の人物関連情報抽出については、人物に関連するすべてのあるいは一部の重要な情報をランキング出力することが一般的であるが、本研究では、人物に関連するラベルの付与を人物ディレクトリの開発を目的として、階層構造を持つ数個の図書館の分類記号を得ることを目的としている。ここに問題設定の新規性がある。Web 上の人物に図書館の分類記号を付与して人物ディレクトリを開発して実験を行った事例は著者らの研究が初めてである。Web ページに図書館の分類記号を付与する研究としては主として英語と米国の分類記号を対象として海外にはいくつかあるが、日本語と日本の図書館の分類記号を対象としたものはほとんどない。先行研究の多くが機械学習手法またはベクトル空間モデルによる照合を行うが、相関索引を用いた間接的な照合を用いる方法はわれわれが調べた限り本研究のみである。6 つの文書の中でタイトルを用いた方法が最も良かったという結果にも新規性がある。これは、Web 上の人物検索という課題において発生する不要な情報を除去するためにタイトル要素の利用が有効であること、ユーザ評価においてはゴミを除去することが重要であることも示唆している。

提案手法の有効性の評価は、情報検索分野における一般的な方法（適合率、再現率、F 値）だけでなく、ユーザによるディレクトリの評価実験により確かめている。提案手法はタイトルから抽出したテキストと相関索引を照合するというシンプルな方法であるが、学習データがない状況でも実現可能で適用範囲が広く実装が容易である。

(2) 履歴書＋地図インタフェース

①提案手法の概要

先行研究[上田 2010]の手法を用いて履歴書を作成する。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する。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、位置情報を付与した学歴と職歴データを作成して地図表示を行う。

以下では、履歴書の作成と、学校と勤務先の抽出について述べる。

②履歴書の作成

先行研究[上田 2010]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。手法は Web からの人物に関する履歴文の抽出、履歴文の分類、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。

本研究では[上田 2010]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。

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。

a 学校の抽出

Step 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。

Step 2．MeCab で履歴文を形態素解析する。

Step 3．学校候補を履歴文から抽出する。

3-a 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。

3-b「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。

Step 4．同義の学校候補のフィルタリング：「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。

b 勤務先の抽出

Step 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。

Step 2．MeCab で履歴文を形態素解析する。

Step 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、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。

Step 4.「名詞」の「固有名詞」の「組織」が存在する場合

4-a「組織」を抽出する。

4-b 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

Step 5.「名詞」の「固有名詞」の「組織」が存在しない場合

5-a「(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。
5-b「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。
5-c「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。

Step 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、学校の抽出とほぼ同様の処理を行う。

④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 2010]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：プロトタイプシステム

⑤評価実験

56 人物のデータセットを用いて評価実験を行った。17 人が政治家、14 人がスポーツ選手、12 人が芸能人、10 人が研究者、1 人が企業家、漫画家、歴史上の人物である。

評価指標には適合率と再現率を用いた。完全一致を 1、部分一致を 2 として、適合率の完全一致を P1、部分一致を P2、再現率の完全一致を R1、部分一誌日を R2 とする。表 1 に学校と勤務先の抽出の実験結果を示す。

表 1：学校と勤務先の抽出

	適合率		再現率	
	P1	P2	R1	R2
学校	91%	97%	84%	90%
	59/65	63/65	59/70	63/70
勤務先	64%	69%	66%	71%
	154/241	166/241	154/233	166/233

完全一致した正解を Google Maps API v3 にかけて最上位の位置情報を評価したところ、学校で 90%(53/59)、勤務先で 53%(81/154)正解であった。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1 件）

① 村上 晴美, 浦 芳伸, 片岡 祐輔, Web 上の人物への図書館の分類記号の付与と人物ディレクトリの開発，システム制御情報学会論文誌，29 巻 2 号，pp．51-64，2016 年，査読有．

〔学会発表〕（計　6 件）

① Zhang Gang, Harumi Murakami, Displaying People with Old Addresses on a Map, AIRS 2015, Brisbane (Australia), pp. 381-386, 2015 年 12 月 2 日，査読有．

② 張 鋼，村上 晴美，昔の住所を持つ人物の地図上への表示，2015 年 9 月 17 日，FIT2016，pp.115-116，愛媛大学（愛媛県松山市），査読無．

③ Harumi Murakami, Chunliang Tang, Suang Wang, Hiroshi Ueda, Vitae and Map Display System for People on the Web, Dalien (Chaina), pp. 348-359, 2014 年 6 月 5 日，査読有．

④ 村上 晴美，個人の人生の記録，2013 年度人工知能学会全国大会（第 23 回），富山国際会議場（富山県富山市），2013 年 6 月 5 日，査読無．

⑤ 唐 春亮，王 爽，上田 洋，村上 晴美，Web 上の人物履歴情報の地図表示システム，2013 年度人工知能学会全国大会(第 27 回)，富山国際会議場(富山県富山市)，2013 年 6 月 5，6 日

⑥ Harumi Murakami, Yoshinobu Ura, Yusuke Kataoka, Assigning Library Classification Numbers to People on the Web, AIRS 2013, Singapore (Singapore), pp. 464-475, 2013 年 12 月 9 日，査読有．

６．研究組織

(1)研究代表者

村上 晴美（MURAKAMI Harumi）
大阪市立大学・大学院創造都市研究科・教授
研究者番号：40305644