様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 25 年 5 月 30 日現在

機関番号：24402
研究種目：基盤研究(C)
研究期間：2010～2012
課題番号：22500219
研究課題名（和文）　ウェブ上の人物要約インタフェースの開発

研究課題名（英文）　Development of an Interface to Summarize 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

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インタフェースの開発である。主要な成果は以下の 2 点である。(1) 人間が Web 上の同姓同名人物を分離する過程を認知科学的に明らかにし、Web 上の同姓同名人物の分離モデルと知識構造モデルを提案した。(2) Web 上の人物に NDC9 を付与する手法を提案して NDC 人物ディレクトリを開発し、評価実験を行って提案手法とディレクトリの有効性を確認した。

研究成果の概要（英文）：The aim of this research is to develop an interface that helps users select people on the Web. The following are its main results: (1) We investigated how humans distinguish people with identical names on the Web by a cognitive science approach and proposed a model for distinguishing individuals and a knowledge-structure model. (2) We presented a method that assigns NDC numbers to people on the Web, developed an NDC-based people search directory, and evaluated the usefulness of our proposed method and system.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,100,000	330,000	1,430,000
2011 年度	800,000	240,000	1,040,000
2012 年度	600,000	180,000	780,000
年度			
年度			
総　計	2,500,000	750,000	3,250,000

研究分野：情報学
科研費の分科・細目：情報学・図書館情報学・人文社会情報学
キーワード：情報検索、Web 人物検索、同姓同名、同姓同名人物の判別モデル、NDC 人物ディレクトリ

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。
代表者はこれまでテキストから人物に関する情報を抽出する研究を行ってきた。本研究は代表者が行ってきた「Web 上の同姓同名人物の識別」に関する研究を発展させるものである。「ユーザによる人物クラスタの選択には職業とキーワードが有用である」との予備実験結果に基づき、職業の要約（生成にあたる）手法の開発を行ってきた。本研究では、人物の識別に有用な情報に関する認知実験を行い、実験結果に基づき、要約手法とインタフェースを開発する。
インタフェースの一つとして、図書館の分類番号である NDC9 を人物に付与して NDC 人物ディレクトリの開発を行う。これは代表者らが行ってきた「Web 情報源を用いた件名と分類の提案」に関する研究を発展させるものである。

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。

３．研究の方法
(1) Web 上の同姓同名人物の分離過程の解明
①　概要
被験者は 14 人（男性 9、女性 5 名、平均年齢 25 歳)である。
先行研究で利用された 20 の日本人の人名を用いて Web 検索を行い、400 件（20 人名×20HTML ファイル）の結果を得た。20 人名には有名人と無名人が混在している。400 件の結果が誰に含まれるのか人手で判定を行った。58 人が存在した。このデータを用いて実験用 Web サイトを開発した。
実験用サイトを用いて一人名毎に 20 件の Web ページを人物に分類させた。一人名毎に二人の被験者をわりあてた。
被験者の分類の過程を、質問紙、プロトコル分析、インタビューを用いて分析した。

②　手続き
人名毎の手続きは以下のとおりである。
a 質問紙調査（該当の人名に関する知識の有無の調査）
b 被験者による 20 ページの分類。発話は記録される。
c 質問紙調査
後述する「識別キーワード」と「特徴キーワード」を記述させる。
被験者が別の人名の実験に参加する場合には、a-c の過程を繰り返す。
一人の被験者につきすべての実験が終了したら最後のインタビュー調査が行われる。
質問紙において「タイトル、スニペット、URL、Web 文書の中から、人物を分離する際に参考になったキーワードを重要度の高いと感じる順に 1～10 個以内で列挙して下さい」と教示して記入させた語を識別キーワードと呼ぶ。「分離された人物に対して、その人物を最も特徴付けると思うキーワードを 1 つ記入して下さい。」と教示して記入させた語を特徴キーワードと呼ぶ。

(2) NDC を用いた人物ディレクトリの開発
20 の日本人の人名を用いて Web 検索を行い、2,000 件（20 人名×100HTML ファイル）の結果を取得し、人手で同姓同名人物に分類したデータセットを利用する。152 人の人物が存在する。
NDC9 の相関索引を用いて人物に NDC9 を付与する手法を提案し、NDC 人物ディレクトリのプロトタイプを開発する。
提案手法の有効性及び開発した NDC 人物ディレクトリの有効性の評価を行う。

４．研究成果
(1) Web 上の同姓同名人物の分離過程の解明
①　識別キーワードと特徴キーワード
記入された識別キーワードの合計は 329、特徴キーワードは 124 であった。
識別キーワードと特徴キーワードを 8 カテゴリ（キーワード、職業、作品、関連する人名、経歴、画像、URL、地名）に分類した。キーワード以外の 7 カテゴリに分類できないものをキーワードカテゴリに分類した。結果を図 1 に示す。

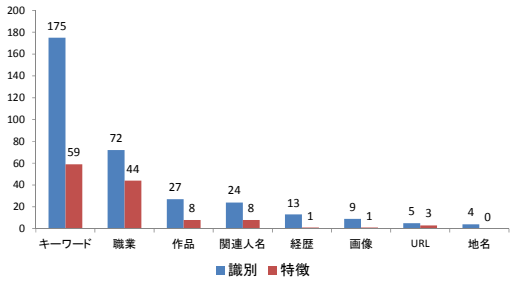

図 1：識別キーワードと特徴キーワード

識別キーワードに関する上位 4 カテゴリは、キーワード（175，53%)、職業（72，22%)、作品（27，8%)、関連する人名（24，7%)の順番であった。
特徴キーワードに関する上位 4 カテゴリはキーワード(59，48%)、職業（44，35%)、作品（8，6%)、関連する人名（8，6%)の順番であった。
以上より、人物を分離するために、キーワード、職業、作品、関連する人名が重要であることがわかる。

② 実在人物と架空人物

データセットには 54 の実在人物（記述は 121 人）と 4 人の架空人物（記述は 7 人）がいた。

識別キーワードと特徴キーワードを実在人物と架空人物に分けて集計した。

実在人物では、作品は該当人物による作品である。架空人物では、作品は該当人物が出現する作品である。関連する人名は、実在人物では実在人物の名前であり、架空人物では架空人物の名前である。

図 2 に実在人物に関する上位 4 カテゴリ、図 3 に架空人物に関する上位 4 カテゴリの分類を示す。

実在人物は全体の結果と類似しているが、架空人物では、作品が職業より重要であることがわかる。

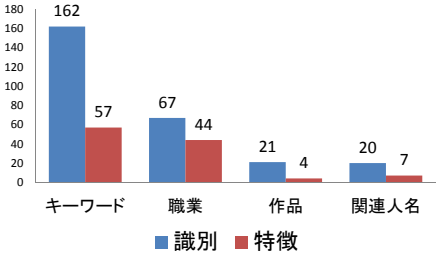

図 2：実在人物の上位 4 カテゴリ

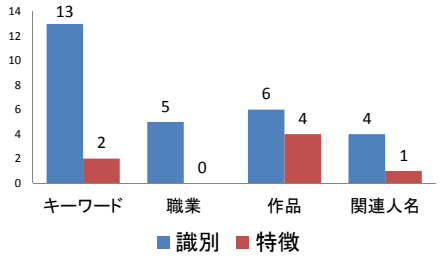

図 3：架空人物の上位 4 カテゴリ

③ プロトコル分析とインタビュー調査

キーワード、職業、作品の重要性は質問紙調査と同じであったが、プロトコル分析とインタビュー調査では、顔画像が重要であることを確認した。また、リンク先を含むサイトが重要であることも発見した。特に人物に関連する情報が少ない場合などに、サイトに関連する情報は重要である。

④ Web 上の同姓同名人物の判別モデル

実験で得られた知見を基に同姓同名人物の判別のモデルを考案した。図 4 では被験者が与えられた人名検索結果一覧を同姓同名人物に分離する過程、図 5 では被験者が人物を判別する際に利用している知識をモデル化した。前者を同姓同名人物の分離モデル、後者を知識構造モデルと呼ぶ。

同姓同名人物の分離モデル（図 4）では、まず、タイトルやスニペットなどを見て、知識構造を参照して、人物が既知か未知かを判別し、既知の場合は人物を分離する。未知の場合や、既知であるが確認が必要な場合はサイトを閲覧する。サイト閲覧により、人物に関連する情報を取得し、人物に関連する知識構造の作成、追加、修正を行う。その後、人物を分離する。

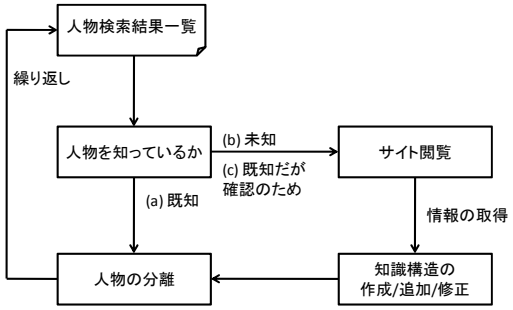

図 4：同姓同名人物の分離モデル

知識構造モデル（図 5）の知識は、人物に関連する知識とサイトに関連する知識に大別できる。

人物に関連する知識は、顔画像と内容テキスト（人物に関連する情報）に分けられる。内容テキストは、実在人物と架空人物に分け、実在人物では、キーワード、職業、作品、経歴、架空人物では、キーワード、作品、職業とした。作品は実在人物の場合該当人物による作品であり、架空人物の場合該当人物の出現する作品である。キーワードの中から関連する人名を抽出して別枠とした。

Web サイトに関連する知識は URL、リンク先のページ、サイトの構造とした。

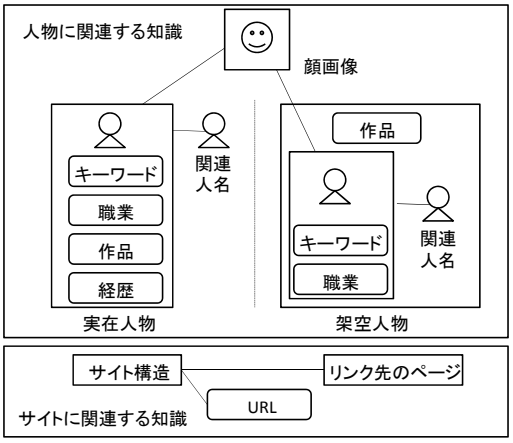

図 5：知識構造モデル

(2) NDC を用いた人物ディレクトリの開発

① 人物に NDC9 を付与する手法

提案手法の主要なアイデアは、NDC9 の相関索引を利用することにある。相関索引は索引語とそれに対応する分類記号の配列である。索引語には分類の細目表に示される名辞をはじめとして必要と判断された用語が含まれている。NDC9 の MRDF 版の相関索引には。分類記号 8,551 件に対して、索引語は 29,514 件存在する。提案手法は、(1) 相関索引語の抽出、(2) NDC の付与、の 2 段階で構成される。

a 相関索引語の抽出

HTML のタグの除去後、相関索引語を抽出する。文字列から複数の索引語が抽出できる場合、最も文字数の多い相関索引語を抽出する。

一字の語と、Web ページで出現頻度が高いもの 100 語程度を不要語として除去する。

b NDC の付与

相関索引語を分類記号に変換する。

以下のとおり人物毎に分類記号 *ndc* のスコアを相対頻度として算出する。

$$score(ndc_i) = \frac{freq(ndc_i)}{\sum_{k=1}^{n} freq(ndc_k)}$$

ただし、n は人物毎の分類記号の異なり数とする。

② プロトタイプ

提案手法とデータセットを用いて人物ディレクトリを試作した。タイトルを用いた提案手法により、上位 5 件の分類記号を人物に付与してから、二次区分以降のカテゴリに割り当てている。

図 6 に二次区分「78 (スポーツ．体育)」の画面例を示す。三次区分の一覧と，二次区分に含まれる人物の一覧が表示されている。

図 6：NDC 人物ディレクトリ

③ 評価実験

a 実験 1

タイトル、全文、スニペット、氏名の前後 50 文字、100 文字、200 文字の 6 種類の文書を比較した。提案手法で付与される最上位の NDC に対して、関連度の 5 段階評価を行った（図 7)。5 を「非常に関連している」、1 を「全く関連していない」とした。タイトルの関連度が最も高かった(3.41)。

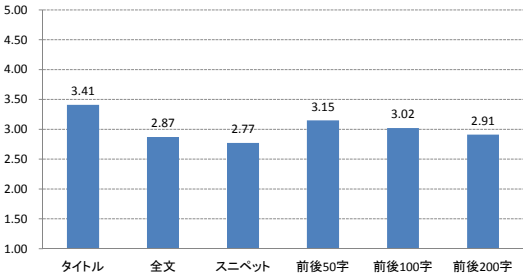

図 7：提案手法の関連度の評価

b 実験 2

6 文書それぞれのプロトタイプを作成し、NDC の二次区分(000-990 の 100 種類)を対象に正解率(正解数/人物数)を求めた（図 8)。タイトルのプロトタイプが最も高かった(39%)。

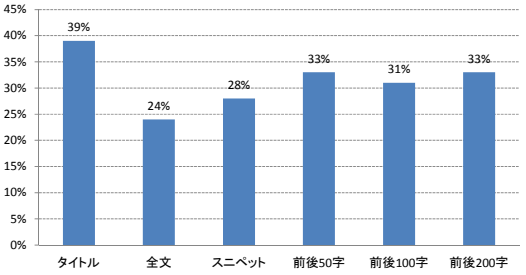

図 8：提案手法の正解率の評価

c 実験 3

3 種類のプロトタイプ(タイトル、全文、前後 50 文字)について被験者 14 人に対してアンケート調査を行った。結果を図 9 に示す。

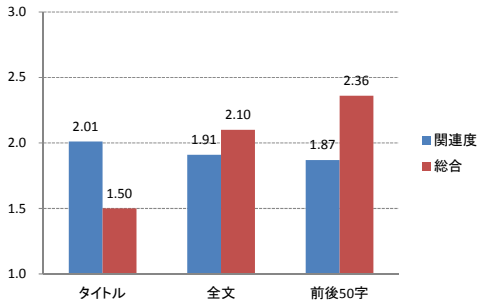

図 9：プロトタイプの評価

NDC の一次区分(000-900 の 10 種類)に登録されている人物の NDC に関して、3 段階評価を行った。3 を「よく当てはまる」、1 を「当てはまらない」とした。タイトルの関連度が最も高かった(2.01)。

試作したプロトタイプの総合的なよさを順位付ける質問でもタイトルが最もよかっ

た(1.50)。タイトルを一位とした被験者からは「無駄な情報が少なくわかりやすい」というコメントがあった。
以上の結果、文書としてWebページのタイトルを使うとよいことがわかった。

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 上田 洋, 村上 晴美, 辰巳 昭治, 著者名典拠作成の自動化を目指して, TP&Dフォーラムシリーズ(整理技術・情報管理等研究論集), 第 20 号, pp.18-34, 2012 年, 査読有.

〔学会発表〕（計 6 件）
① 片岡 祐輔, 浦 芳伸, 村上 晴美, NDCを用いた人物ディレクトリの評価実験, 電子情報通信学会 2013年総合大会 情報・システムソサイエティ特別企画 学生ポスターセッション予稿集, pp.34, 岐阜大学, 2013年3月19日, 査読無.

② Harumi Murakami and Yuki Miyake, How Do Humans Distinguish Different People with Identical Names on the Web?: A Cognitive Science Approach, Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012), pp. 2475-2478, Maui, USA, 2012年10月31日, 査読有.

③ 王 爽, 浦 芳伸, 上田 洋, 村上 晴美, Web上の人物履歴情報の地図上への表示, 電子情報通信学会 2012 年総合大会 情報・システムソサイエティ特別企画 学生ポスターセッション予稿集, pp.99, 岡山大学, 2012年3月21日, 査読無.

④ Harumi Murakami and Yoshinobu Ura, People Search using NDC Classification System, Proceedings of the CIKM 2011 4th Workshop on Exploiting Semantic Annotation in Information Retrieval, Glasgow, UK, 2011年10月28日, 査読有.

⑤ 三宅 悠生, 村上 晴美, 人はWeb上の同姓同名人物をどのように判別しているのか, 電子情報通信学会第二種研究会資料(第19回Webインテリジェンスとインタラクション研究会), pp.73-76, 学術総合センター, 2011年3月8日, 査読無.

⑥ 浦 芳伸, 村上 晴美, NDC を用いた人物ディレクトリの開発, 情報処理学会第73回全国大会講演論文集, Vol.1, pp.651-652, 2011年3月4日, 査読無.

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

(2) 研究分担者
辰巳　昭治（TATSUMI Shoji）
大阪市立大学・大学院工学研究科・教授
研究者番号：80124733