様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 １７ 日現在

機関番号：２４４０２
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００４４０

研究課題名（和文）Web上の人物の概要の作成

研究課題名（英文）Creating Summary of People on the Web

研究代表者
村上　晴美（Murakami, Harumi）

大阪市立大学・大学院工学研究科・教授

研究者番号：４０３０５６４４

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：研究の全体構想は「Web上の人物を選択するためのインタフェースの開発」であり、研究の目的は「Web上の人物を要約するための手法の開発」である。80人物のデータセットを用いた主要な成果は以下の2点である。(1) Web上の人物にWikipediaの第一文風の概要文を作成する手法を提案した。提案手法は属性情報（氏名のよみ、生年月日、没年月日、出身地、職業、所属、役職）抽出と概要文作成からなる。(2) Web上の人物に件名（NDLSH）を付与する手法を検討した。(a) 検索エンジンランキング、(b) 文書内の位置、(c) 同義語、(d) 文書頻度の組合せを調査した。

研究成果の学術的意義や社会的意義
本研究では、Web上の人物にWikipediaの第一文風の概要文を作成する手法を提案し、また、Web上の人物に件名を与える手法を検討した。Web上の文書の概要を作成することやWeb上の文書に件名を付与する研究は存在するが、Web上の人物に焦点をあててWikipediaの第一文風の概要文を作成する研究や、件名(NDLSH）を付与する研究は代表者の知る限り初めてであり、そこに学術的・社会的意義がある。

研究成果の概要（英文）：The overall concept of this research is to develop interfaces that help users select people on the Web. The aim of this research is to develop methodology of summarizing people on the web. The following are its main results using an 80-person dataset. (1) We developed a method of generating summary sentences for the results of web people search. We extracted attribute information about people (furigana reading for person name, birth date, death date, place of birth, vocation, organization, and position) and generated summary sentences whose style resembles the first sentence of Wikipedia. (2) We investigated a method that assigns National Diet Subject Headings (NDLSH) to the results of web people search. We examined the following combination of factors: (a) web-page rank, (b) position inside HTML, (c) synonyms, and (d) document frequency.

研究分野：情報検索

キーワード：情報検索　Web人物検索　概要文　件名

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

人物検索は情報検索において重要な課題の一つである。Web 上の人名検索においては、人名の曖昧性解消が重要な課題である。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 上の人物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，または付与する（本研究では「要約」と呼ぶ）」。

本研究は科学研究費（No. 22500219）（No. 25330385）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」「NDC を用いた人物ディレクトリの開発」「履歴書＋地図インタフェースの開発」であった。

２．研究の目的

研究の全体構想は「Web 上の人物を選択するためのインタフェースの開発」であり、「Web 上の人物を要約するための手法の開発」を目的とする。具体的には人物の「要約」手法と 2 種類のインタフェース（表と概要文）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) Wikipedia の第一文風の概要文の作成、(2) 件名の付与について述べる。

３．研究の方法

(1) データセットの作成

先行研究[佐藤 05]による 20 の日本人氏名を用いて、Google Web APIs を用いて検索を行い、50 件の検索結果を取得し、人手で人物毎に分離してデータセットを作成した。データセットには 80 人物存在する。以下ではこのデータセットを用いた。

(2) Wikipedia の第一文風の概要文の作成

データセット中には Wikipedia のページがある人物が 14 人物存在した。実在人物が 12 人、架空人物が 2 人であり、実在人物の中生存人物が 8 人であった。この実在人物 12 人の第一文の先頭のフォーマットは概ね以下のとおりであることを確認した。

氏 名（氏のよみ 名のよみ、生年月日 － 没年月日）は、地域等の職業等。

地域等は 8 人が「日本」であり、2 人が「東京都出身」と「福島県出身」、1 人が「アメリカ合衆国を拠点に活躍する大阪府枚方市出身の日本」であり、1 人は存在しなかった。職業等は、1-5 の職業（所属等を含む）が列挙されており、平均は 1。9、内訳は職業が 16、所属＋役職が 6、学位が 1 であった。なお、12 人物中 11 人物に第二文以降があり、第二文の内容は所属＋役職が 9、他が 5（本名、旧姓、出身地、血液型、業績）であった。

上記の分析により、本研究では、氏のよみ、名のよみ、生年月日、没年月日、出身地、職業、所属と役職を抽出対象とすることにした。研究の手法は、属性情報抽出と概要文生成に大別できる。

① 属性情報抽出

[氏名のよみ]

氏と名で分けて処理を行い後で結合する。前処理として文書のカタカナをひらがなに、大文字を小文字に変換する。

「人名漢字辞典‐読み方検索（http://kanji。reader。bz/）」を用いてよみ候補を作成し、小文字のローマ字表記を加える。（例）伊庭⇒いば、iba；幸人⇒ゆきと、ゆきひと、yukito、yukihito

よみ候補で文書を検索し、一致したものを抽出する。複数ある場合は前方（Web の上位⇒文書の上方）を優先する。

なお、複数ある場合は以下同様（前方優先）である。

[生年月日と没年月日]

「生まれ」や「年月日」等に着目した正規表現を用いて西暦の生年月日及び没年月日を抽出する。

[出身地]

人物のプロフィールが含まれやすい氏名の前後 100 文字の文字列を取得する。その中から「出身」の前後 10 文字ずつ、「生まれ」の前 10 文字にある「都道府県」を都道府県辞書を用いて抽

出した。その際「出身者」等の表記のある箇所を除外した。

[職業]
Wikipedia の職業一覧ページの最も長い職業名を参考にして、氏名の前後の 20 文字ずつを取得する。その中から「師」「士」等や、最後が「ー」で終わる 4 文字以上のカタカナが含まれている文字列を抽出し、形態素解析をかけ、連続した名詞を結合する。結合された文字列の内、最後の文字が「師」「士」「ー」等となっている文字列を職業として抽出する。

[所属と役職]
所属と役職に分けて抽出し、後で結合する。

[所属]
組織を表す文字列は非常に多いため上位 5 件の文書を用いる。
「センター」「営業所」「病院」「大学」等の文字列が含まれている行を抽出し、形態素解析をかけ、連続した名詞を結合する。この際、経歴を表す「入学」「卒業」等を除外する。結合された文字列の内、語尾が抽出に用いた文字列になっているものを所属として抽出する。
また、「株式会社」「クリニック」「スタジオ」等の固有名詞が連続すると考えられる文字列では、連続した記号以外の形態素を結合する。結合された文字列の内、文字列の先頭もしくは語尾が抽出に用いた語となっているものを所属として抽出する。

[役職]
所属を抽出できた場合のみ、 抽出された所属名を用いて役職を抽出する。
上位 5 件に出現する所属名の後方 50 文字から、「社長」「所長」「院長」「取締役」「幹事」等が含まれている文字列を抽出し、前方に連続した名詞を結合する。

② 概要文生成
以下に、概要文生成のアルゴリズムを示す。人物の属性情報を与えると、「氏 名（氏のよみ 名のよみ、生年月日 － 没年月日）は、出身地の職業。所属役職。」という概要文を生成する。

```
function Generate-Japanese-Summary(person-attribute-information)
returns a summary
    Name ← 氏 名
    Yomi ← 氏のよみ　名のよみ
    Birth ← 生年月日
    Death ← 没年月日
    Birthplace ← 出身地
    Vocation ← 職業
    Organization ← 所属
    Position ← 役職
    if Birthplace is empty then Birthplace ← “ 日本の ”
    else Birthplace ← Birthplace + “ 出身の ”
    end if
    if Vocation is empty then Vocation ← “ 人物。”
    else Vocation ← Vocation + “ 。”
    end if
    if Organization is not empty then
      if Position is empty then Position ← “ 所属。”
      else Position ← Position + “ 。”
      end if
    end if
    Summary ← Name + “（”+ Yomi + “、”+ Birth + “ － ”+ Death + “）は、”
    + Birthplace + Vocation + Organization + Position
    return Summary
```

(3) 件名の付与
データセットに NDLSH を付与する。
まず、NDLSH の標目と同義語を抽出する。この時、標目からは半角英数字 2 文字以下、全角 1 文字のみ、――（ハイフン 2 つ）が含まれる語はあまり重要ではないあるいは照合がうまくいかないと考えて除去している。
標目と同義語について、文字列が長い方がより詳細な意味を付与できると考えて、文字列の長いものから順に、以下の(a)と(b)で与えられるタグを除いた HTML 文書と照合してカウントする。一致した箇所は半角空白 1 つに置き換えて、次の標目または同義語を処理する。たとえば文書中の「人工知能」という文字列を処理する際に標目「人工知能」はカウントされるが標目

「知能」はカウントされない。標目や同義語をカウントした後に重み付けを行い該当する標目のスコアを算出する。

組合せ条件として以下の 4 種類を用意した。

(a) Web ページの検索ランキングの利用：人物毎の上位 1、 3、 5、 10 件および全件の 5 パターン。

(b) HTML 文書内の位置の利用：タイトル、全文、検索語（人名）の前後の文字（前後 20、 40、 60、 80、 100、 150、 200）の 9 パターン

(c) 同義語の利用 ：同義語を利用しない、 標目の 0。5 倍の重みで利用、 標目と同じ重みで利用の 3 パターン

(d) 標目および同義語の文書頻度の利用：何もしない、文書頻度（df）/利用した全文書数（N）をかける、利用した全文書数（N）/文書頻度（df）をかけるの 3 パターン

これらを組み合わせると 5×9×3×3=405 パターンとなる。図 1 に標目のスコア計算例を示す。最上位のスコアを持つ標目を該当人物に付与する。ない場合は「なし」とする。

４．研究成果

(1) Wikipedia の第一文風の概要文の作成

図 1 に、人物「三浦麻子 0」のクラスタ（HTML32 個）を入力とした Google のナレッジグラフ風の出力結果を示す。本研究の手法で出力される氏名のよみ、職業、概要文から構成される。

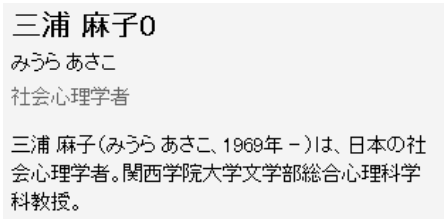

図 1：出力例

概要文は 100%(80/80)生成できた。ただし、15%にあたる 12 人物は「氏名は、日本の人物。」というほぼ無意味な結果である。この中 92%(11/12)にあたる 11 人物ではページ数が 1 であった。

属性情報抽出については、適合率、再現率で評価した。表 1 に属性情報抽出の結果を示す。氏名のよみ、生年月日、没年月日、出身地の適合率は 80%以上と比較的良好であったが、職業と所属と役職の適合率は 60%台であり改良の余地が大きい。再現率も職業、所属と役職では低く改善の必要がある。

表１：属性情報抽出結果

氏名のよみ		生年月日		没年月日		出身地	
適合率	再現率	適合率	再現率	適合率	再現率	適合率	再現率
100%	95%	82%	70%	100%	75%	93%	93%
(37/37)	(35/37)	(14/17)	(14/20)	(3/3)	(3/4)	(13/14)	(13/14)
職業		所属		役職			
適合率	再現率	適合率	再現率	適合率	再現率		
67%	24%	62%	59%	66%	38%		
(16/24)	(16/67)	(39/63)	(39/66)	(19/29)	(19/50)		

(2) 件名の付与

正解率においてどのパターンが最も良かったかを表 2 に示す。

全体（80 人物）について最も良かったパターンは、「上位 10 件、人名の前後 100 文字（合計 200 文字）、同義語 0.5 倍、df/N」であり、正解率は 26.3%（21/80）であった。

人物毎の文書数によって傾向が異なることが観察されたため、1、2、3 文書以上、11 文書以上、3 文書以上 10 文書以下に分けて調べた。1 文書しかない 35 人物の場合「全文（同義語は利用しない）」が最も良かった。3 文書以上の 33 人物の場合全体とほぼ同じ（違いは同義語の倍率のみ）であり、11 文書以上の 8 人物と同じ結果であった。

全体として、Web ページ全てよりも上位 10 件に抑え、HTML 文書の全文よりも人名の前後の文字列を利用し、同義語を利用し、多くの文書に出現する語に重み付けすることがよかった。ただし、1 文書しかない人物については全文から標目をカウントするだけが最も良かった。

表 2：正解率の良かったパターン

文書数	上位件数	利用箇所	同義語	文書頻度	正解率
全体	10	前後 100	0.5	df/N	0.263 (21/80)
1	1	全文	0	1	0.286 (10/35)
2	3	前後 200	0.5	1	0.333 (4/12)
3 以上	10	前後 100	1	df/N	0.364 (12/33)
11 以上	10	前後 100	1	df/N	0.500 (9/18)
3〜10	5	前後 60	0.5	df/N	0.267 (4/15)

５．主な発表論文等

〔雑誌論文〕（計２件）

① Masayuki Shimokura, Harumi Murakami, Assigning NDLSH Headings to People on the Web, Tseng YH. et al. (eds) Information Retrieval Technology. AIRS 2018. Lecture Notes in Computer Science, 査読有, vol 11292, 2018, pp. 189-195. DOI: 10.1007/978-3-030-03520-4_18

② Harumi Murakami, Toshimune Konishi, Yoshinobu Ura, Generating Wikipedia-Like Biographical Sentences from Web People Search Results, 2017 6th IIAI International Congress on Advanced Applied Informatics IIAI-AAI 2017, 査読有, 2017, pp.992-993. DOI:10.1109/IIAI-AAI.2017.38

〔学会発表〕（計５件）

① 下倉 雅行, 村上 晴美, Web 上の人物への BSH の付与, 2018 年度人工知能学会全国大会（第 32 回）, 2018.

② Masayuki Shimokura, Harumi Murakami, Assigning NDLSH Headings to People on the Web, AIRS 2017, 2017.

③ Harumi Murakami, Toshimune Konishi, Yoshinobu Ura, Generating Wikipedia-Like Biographical Sentences from Web People Search Results, IIAI-AAI 2017, 2017.

④ 下倉 雅行, 村上 晴美, Web 上の人物への NDLSH の付与, 2017 年度人工知能学会全国大会 (第 31 回), 2017.

⑤ 村上 晴美, 小西 利宗, 浦 芳伸, Web 上の人物の概要文の作成, 2016 年度人工知能学会全国大会(第 30 回), 2016.

〔その他〕
ホームページ等
http://murakami.media.osaka-cu.ac.jp/research/WebPeople/

６．研究組織

(1)研究分担者
なし