論文要旨

Web 上の人名検索においては，検索結果における同姓同名人物の識別が重要な課題となってきている．本研究の目的は，人名検索の結果取得された Web ページを同姓同名人物毎に識別・分類したクラスタ（人物クラスタ）に，人物の識別に有用なラベルを付与することである．本研究では，人物クラスタに一つの「職業関連情報」をラベルとして付与する手法を提案する.「職業関連情報」とは，厳密に職業と定義される語だけではなく，幅広く職業と考えられる語や，職業の推定に有用と思われる語も含めた，職業に関連する情報である．提案手法は，(a) HTML 構造と簡単なヒューリスティックを用いた職業関連情報候補抽出，(b) 出現頻度，同義クラスタ作成，Web 検索エンジンを用いた職業関連情報作成から構成される．実験の結果，提案手法の有効性を確認した．

論　文

Web上の同姓同名人物識別のための職業関連情報の抽出*

上田　洋†・村上　晴美‡・辰巳　昭治†

Extracting Vocation-Related Information for Distinguishing Different People with Identical Names on the Web *

Hiroshi Ueda†, Harumi Murakami‡ and Shoji Tatsumi†

Distinguishing different people with identical names is becoming more and more important in person searches on the Web. The aim of this research is to dispatch useful labels for identifying persons in “ person clusters, ” which are generated as a result of person searches on the Web. In this paper, we propose a method to label person clusters with “ vocation-related information. ” The vocation-related information includes broader terms that may be considered as vocations, and terms that are useful to infer vocations, not only those rigorously defined as vocations. Our method is based on (a) extracting candidates of vocation-related information by using HTML structures and simple heuristics, and (b) generating vocation-related information by using term frequencies, synonym clustering, and Web search engines. Experimental results revealed the usefulness of the proposed method.

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．一方, [1] によれば，Web検索におけるクエリの約3割は人名を含むとされる．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以前にもまして重要になってきていると考える．

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, [2–5] 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ページを一つずつ閲覧しなければならず，ユーザにとって非常に負荷が高い．

本研究では，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当なラベルを付与することを検討する．ユーザは，ラベルを参考にすることにより，クラスタ内のWebページを閲覧することなくクラスタを選択できる．代表的な先行研究としてはWanらの研究 [6] がある．Wanらは，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一つラベル付けを行っている．ラベルは「Title（以下，肩書）」が取得できた場合には肩書，取得できなかった場合には代替情報を付与している．たとえば，人名「David Lee」に対して「David Lee Roth, Roth Tickts(32)」「David Lee Murphy, Artist (12)」「David Lee Smith, Fan Sites(6)」「David M. Lee, Professor (5)」などが表示される[1]．肩書あるいは代替情報をクラスタに付与するアイディアは非常に有用であるが，以下のような課題がある． David Lee Rothは著名なミュージシャンであるが，代替情報の「ロスのチケット」より職業である「ミュージシャン」を付与したほうが識別が容易なので

* 原稿受付　2008年7月14日

† 大阪市立大学　大学院　工学研究科　Graduate School of Engineering, Osaka City University; 3-3-138 Sugimoto-cho, Sumiyoshiku, Osaka city, Osaka 558-8585, JAPAN

‡ 大阪市立大学　大学院　創造都市研究科　Graduate School for Creative Cites, Osaka City University; 3-3-138 Sugimotocho, Sumiyoshiku, Osaka city, Osaka 558-8585, JAPAN

Key Words: Vocation-Related Information, Information Extraction, Web Person Search, Person Name Disambiguation

[1] () 内はWebページ数である．「肩書」として一部「職業」も抽出している．

はないだろうか．また，代替情報の「ロスのチケット」「ファンサイト」は，語の属性が雑多であり，クラスタを一瞥では比較しにくい．さらに，Web 上の同姓同名人物に「教授」などの肩書の人物が複数いることがあるが，これだけでは識別できない．

我々は，人物毎のクラスタを一瞥で比較するためには，種類が統一された属性情報をラベルとして付与することが有用であると考える．本研究では，人物識別に有用な属性情報として，職業に注目する．ただし，職業とは何か，言葉の定義は人によって異なる．たとえば，第 2，3 著者の職業は研究者，工学研究者，教育者，大学教員，教授，大学教授などが考えられるが，この中のどれをラベルとすればよいだろうか．仮に「教授」を採用した場合，先述のように，検索結果に教授が複数いることを考慮すると「所属名」もつけたほうがよいかもしれない．

このような背景から，本研究では，厳密に職業と定義される語だけではなく，幅広く職業と考えられる語や，職業の推定に有用と思われる語も含めた，「職業に関連する情報」を「職業関連情報」としてラベル付けに用いる．

本研究では，人物毎に分けられた Web ページクラスタに一つの職業関連情報をラベルとして付与する手法を提案する．本研究における職業関連情報とは，(1) 職業を表す語，(2) 所属と役職を表す語，(3) 著作と役割を表す語である．

本論文の構成は以下のとおりである．まず，2 章で Web から抽出可能な職業関連情報について論じ，3 章では提案手法について述べる．4 章では，提案手法で実際に得られる職業関連情報の例を示し，5 章で，提案手法の有効性を確認するために行った実験について述べる．6 章で関連研究と議論を述べ，最後に 7 章にてまとめを行う．

2. Web 上の人物に付与する職業関連情報

2.1 職業とは

職業とは何か．広辞苑 [7] によると職業とは「日常従事する業務．生計を立てるための仕事」である．「労働省編職業分類 [8]」によると「職務の内容である仕事や課せられた責任を遂行するために要求されている技能，知識，能力などの共通性または類似性によってまとめられた一群の職務」である．

さて，職業分類のページ[1] によると，以下のような例があげられている．子どもに将来なりたい職業を聞くと，「学校の先生」「医者」「野球選手」などがあげられる．大人が何かに申し込みをするときの職業欄には，「会社員」「公務員」「税理士」などがある．求人誌上の募集職種として，「販売スタッフ」「一般事務」「個別指導教師」などが記載されている．この中で，どれが職業でどれが職業でないのであろうか．

このように，日常生活で使われる職業の例は雑多である．何を職業と呼ぶかは人によって異なる．本研究においてクラスタに付与するラベルはどのようなものを選ぶべきであろうか．

[1] http://homepage3.nifty.com/54321/syokugyou.html

Table 1 Examples of Vocations Contained in Subtyping Classification in [8]

ペレット厚さ測定工（半導体製品製造）
ペレットエッチング工（半導体製品製造）
ペレット加工工（半導体製品製造）
ペレット工
ペレット工（飼料製造）
ペレット製造工（金属製錬）
ペレットソルダーリング工
ベロアー織工
ベローズ研磨工（車輪製造）
ベロー製造工
変圧運転員（変電所）
変圧器組立工
変圧器組立工（電気通信機）
変圧器組立工（電子機器用）
変圧器仕上工
変圧器修理工
変圧器製造技術者
変圧器・変流器・変成器組立・調整工
変圧電器技術者
ペン画工
弁柄製造工
ペンキ画工
ペンキ職
ペンキ職人
ペンキ塗装工
編曲家
編曲者
偏光板切断工（プラスチック）
編光器仕上工
弁護士

2.2 辞書中の語か Web ページ中の語か

職業として，辞書（シソーラスや分類表などを含む）に含まれる語を提示するか，Web ページに含まれる語を提示するかの選択肢がある．

まずは辞書の利用を検討した．先にあげた労働省編職業分類は日本における代表的な職業分類の一つである．労働省編職業分類では，業種別に職業が分類され，大分類，中分類，小分類，細分類の 4 階層から構成される．大分類には 9 個，中分類 80 個，小分類 379 個，細分類 2167 個の職業がある．Table 1 に，細分類の例の中，五十音順の索引に記載されている「ペレット厚さ測定工（半導体製品製造）」から「弁護士」までを示す．一瞥してわかるように，日常生活では使われない語が多く記載されている．また，比較的新しい職業は記載されていない．このため，この分類を本研究の目的に利用するのは不向きであると考えた．

一方，Web 上では人々による辞典や辞書の作成がさかんである．代表的な例として Wikipedia があり，職業辞書[1] が作成されつつある．我々も予備研究において Wikipedia の辞書を利用しようと試みた [9] が，職業数が少なすぎる (2007 年 4 月当時，203 の職業)，職業の記載に偏りがある，などの理由により断念した．

他の辞書の利用も検討したが，上記の理由のどれかにあてはまった．そのため，本研究では，Web ページに含まれる語から職業を抽出して提示する方針とした．

2.3 Web ページからの職業抽出と人物へのラベル付け

Web ページからの職業抽出に関しては，たとえば，「語尾に「士」がつく名詞を抽出する」といったヒューリスティックを用いて一部の専門的な職業（例：弁護士）を抽出することは容易である．

しかし，Web ページには人物の職業が必ずしも明示的に記述されるわけではない．特に専門的な職業でないほどその傾向は強い．たとえば，「会社員」は職業として一般的であると考えられるが，Web 上で「会社員」を名乗ったり，「会社員」として言及される人物はほとんど見かけない．また，日本では，職業を聞かれて「事務員」「販売スタッフ」などではなく「会社名」を答えることが多いことが知られており，所属とその所属での役職が広く職業として扱われている．

また，人物へのラベル付けの問題としては「複数の職業候補からどれを選択すべきか規範がない」ことがあげられる．職業の定義や日常生活での使われ方は多様であり，何を職業とするか人によって異なる．また，人物の識別のためには，職業である語「教員」や「研究者」よりも肩書である「教授」の方が有用な場合があるだろうし，その場合は，組織名も一緒に提示するほうが有用であるだろう．さらには，ある研究者が，学問分野によって著名な場合は「○○学者」，学問分野よりも所属が知られている場合は「○○大学○学部教授」のように提示するほうが有用と考えられる．

このような背景から，本研究では職業関連情報をラベル付けに用いることにした．また，Web ページに含まれる複数の職業関連情報の中から，人物の識別に有用な職業関連情報を一つ付与することを目指す．

厳密には職業を表す語ではないが職業に関連する情報として，所属名と役職をあげたが，そのほかにも，たとえば，著作と役割を表す語が考えられる．有名な作家であれば，Web 上で「作家」や「小説家」と記述されることが多いが，無名な作家の場合では記述されない．このような場合，著作のタイトルとその役割（作者，編集者など）を提示することが有用であると考える．

以上の理由により，本研究における職業関連情報として，(1) 職業を表す語，(2) 所属と役職を表す語，(3) 著作と役割を表す語を用いることとした．

[1] http://ja.wikipedia.org/wiki/%E8%81%B7%E6%A5%AD%E4%B8%80%E8%A6%A7

3. 提案手法

本研究では，人物毎に分けられた Web ページクラスタ (以下，人物クラスタ) に一つの職業関連情報をラベルとして付与する手法を提案する．

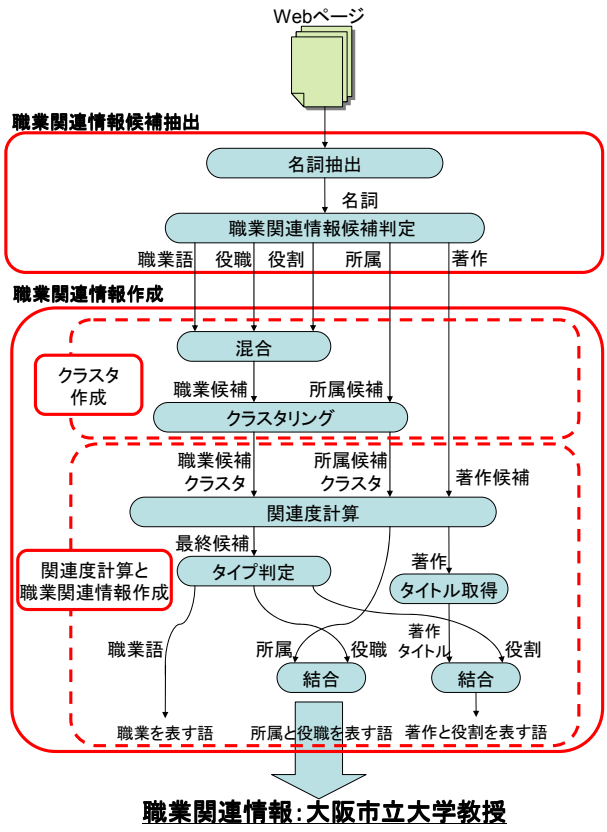


Fig. 1 Overview of our Approach

3.1 提案手法概要

人物毎に分けられた Web ページ中の語を用いて，職業関連情報を一つ作成する．職業関連情報とは，(1) 職業を表す語，(2) 所属と役職を表す語，(3) 著作と役割を表す語である．提案手法の概要を Fig. 1 に示す．

提案手法は，職業関連情報候補抽出処理と職業関連情報作成処理に大別される．

職業関連情報候補抽出処理では，Web ページの中から HTML 構造に着目して人物に関連する名詞を抽出し，その名詞からヒューリスティックを用いて職業関連情報の候補（職業語候補，役職候補，役割候補，所属候補，著作候補）を抽出する．

職業関連情報作成処理では，職業関連情報の候補について，著作候補を除いてクラスタ（職業候補クラスタと所属候補クラスタ）作成を行った後，職業関連情報を作成する．まず，職業候補クラスタについて，出現頻度と Web 検索エンジンを用いた関連度計算を行い最終候補を出力する．次に，最終候補のタイプ判定を行い，(1) 職

業を表す語である場合はそれを職業関連情報とし，(2) 役職を表す語である場合は，所属候補クラスタを用いて，出現頻度と Web 検索エンジンを用いた関連度計算に基づき所属を出力し，役職と組み合わせて職業関連情報とし，(3) 役割を表す語である場合には，文字距離を用いた関連度計算に基づき著作を出力し，役割と組み合わせて提示する．

3.2 職業関連情報候補抽出

職業関連情報候補抽出処理では，Web ページの中から HTML 構造に着目して人物に関連する名詞を抽出し，その名詞からヒューリスティックを用いて職業関連情報の候補（職業語候補，役職候補，役割候補，所属候補，著作候補）を抽出する．

3.2.1 名詞抽出

人名で検索した結果取得される Web ページ全体には，該当人物に関連する情報と関連しない情報が含まれる．提案手法では，該当人物に関連しない情報を排除し，関連する情報だけをできるだけ多く抽出することを目指す．

まず，HTML 構造に着目し，以下の 5 つの要素内に含まれる部分から，形態素解析を用いて名詞を抽出する．

1. 人名検索にて用いた氏名（以下，検索氏名）が出現する p 要素（段落) 内
2. 検索氏名が出現する tr 要素（表の列) 内
3. 検索氏名が 1 行目または 2 行目に出現する table 要素 (表）内
4. 検索氏名が出現する title 要素内
5. 検索氏名が出現する h1〜h3 要素（見出し）内

次に，検索氏名の位置関係に着目し，以下の名詞を抽出する．

6. 検索氏名の直後に出現する丸括弧内の名詞
7. 検索氏名の直前と直後に出現する名詞

6，7 は，HTML で記述されていない，プレーンテキストや PDF ファイルなどでも適用可能である．

本研究で抽出する名詞とは，単名詞，または，連接する名詞やアルファベットや未知語をつなげた複合名詞とする．形態素解析には ChaSen [1] を用いる．

最後に，次節における著作候補の抽出のため，正規表現により ISBN 番号を抽出し，名詞に加えておく．

3.2.2 職業関連情報候補判定

得られた名詞について，ヒューリスティックを用いて職業関連情報の候補かどうかを判定する．

判定は，(1) 職業を表す語，(3) 著作と役割を表す語，(2) 所属と役職を表す語の順番で行う．以下では判定方法について述べる．

(1) 職業を表す語

名詞の語尾に着目した 17 種類のヒューリスティックを選定した．いずれかに合致する名詞を，職業を表す語（以下，職業語）の候補（以下，職業語候補）とする．選定したヒューリスティックと判定される職業語候補の例を以下に示す．

- 「士」で終わる名詞
  （例：弁護士，弁理士，税理士）
- 「優」で終わる名詞
  （例：俳優，女優，声優）
- カタカナ 4 文字以上で構成され，かつ語尾が「ー」で終わる名詞
  （例：アナウンサー，レーサー）
- カタカナ 4 文字以上で構成され，かつ語尾が「スト」で終わる名詞
  （例：ピアニスト，ギタリスト）

(2) 所属と役職を表す語

所属と役職の 2 つに分けて判定する．

所属を表す語の候補（以下，所属候補）の判定方法は 2 種類考案した．

まず，固有表現抽出システム NExT[10] に付属する組織名辞書を利用して判定する．組織名辞書には,「大学」,「病院」,「株式会社」などが含まれる．以下に判定方法と所属候補「大阪市立大学」の判定例を Fig. 2 に示す．

- 「NExT の組織名辞書に含まれる 2 文字以上の語」が語尾につき，かつ,「合致した組織名辞書の語」の文字数+2 以上の文字数の名詞

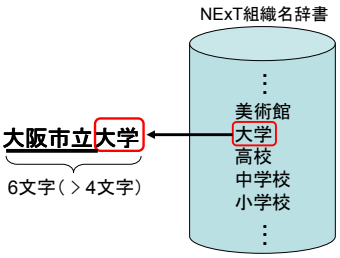


Fig. 2 Judging Organization Term Candidates

次に，14 種類のヒューリスティックを選定した．以下に例を示す．

- 「(株)」で始まり，かつ 5 文字以上で構成される名詞
- 「(有)」で始まり，かつ 5 文字以上で構成される名詞
- 「株式会社」で始まり，かつ 5 文字以上で構成される名詞

役職に関しては，(1) と同じように，名詞の語尾に着目し，15 種類のヒューリスティックを選定し，合致するものを役職候補とする．以下に例を示す．

- 「員」で終わる名詞
- 「長」で終わる名詞
- 「者」で終わる名詞
- 「教授」で終わる名詞

(3) 著作と役割を表す語

著作と役割の 2 つに分けて判定する．

著作については，前節で抽出した ISBN 番号をそのまま著作候補とする．

[1] http://chasen-legacy.sourceforge.jp/

役割については，以下の 3 つのヒューリスティックを用いて判定し，役割候補とする．

- 「著者」で終わる名詞
- 「編者」で終わる名詞
- 「訳者」で終わる名詞

3.3 職業関連情報作成

職業関連情報の候補（職業語候補，所属候補，役職候補，著作候補，役割候補）から，職業関連情報を作成する．

著作候補を除いてクラスタ（職業候補クラスタと所属候補クラスタ）作成を行った後，職業関連情報を作成する．

3.3.1 クラスタ作成

表記のゆれを吸収するために，語尾に着目したクラスタリングを行い，著作候補を除いて同義のクラスタを作成する．

職業語候補，所属候補，役職候補（以下，この 3 つの候補を職業候補と呼ぶ）の 3 つを混ぜて職業候補クラスタを作成し，所属候補から所属候補クラスタを作成する．

職業候補のクラスタリング手法を Fig. 3 に示す．

提案手法により，たとえば，第 3 著者の 5 つの役職候補である「教授」「大阪市大教授」「大阪市立大学教授」「大阪市立大学大学院教授」「大阪市立大学大学院工学研究科教授」を 1 つの職業候補クラスタにまとめることができる．

Step.1 候補の出現頻度を計算する．得られた候補と計算の結果得られた出現頻度の組を持つリスト（以下，候補リスト）を作成する．

Step.2 候補リストのうち，構成文字数が最も少ないものを選択し，その候補と出現頻度の組のみを含むクラスタを作成する．

Step.3 クラスタに属さない全て候補の語尾と，**Step.2** で選択した候補をパターンマッチにより比較する．**Step.2** で選択した候補が含まれれば **Step.2** で作成したクラスタに追加する．

Step.4 クラスタに追加した候補を候補リストから削除する．

Step.5 候補リストの中の候補がなくなるまで，**Step.2** から **Step.4** を実行する．

Fig. 3 Generating Vocation Candidate Clusters

所属候補のクラスタリング手法は，職業候補クラスタ作成手法（Fig. 3 ）の **Step.3** のみが異なる（Fig. 4 ）．

Step.3 クラスタに属さない全て候補の語頭と，**Step.2** で選択した候補をパターンマッチにより比較する．**Step.2** で選択した候補が含まれれば **Step.2** で作成したクラスタに追加する．

Fig. 4 **Step.3** of Generating Organization Candidate Clusters

3.3.2 関連度計算と職業関連情報作成

職業候補クラスタから，クラスタ内に含まれる候補の出現頻度の合計が最も多いクラスタを選択する．最頻度が複数存在する場合は，複数選択する．

選択されたクラスタの中に含まれる候補が 1 つのみの場合，その候補を最終候補とする．

候補が複数だった場合（複数候補と呼ぶ）に 1 つを選択する必要がある．複数から 1 つ選択する手法として，出現頻度や単語距離，TF・IDF など用いる手法が存在する．これらの手法は，扱うデータ内部の情報（以下，内部情報）を元にスコア計算を行う．内部情報を元にしたスコア計算は，データが少ない場合や情報に偏りがある場合にうまくいかないことがある．そこで提案手法では，データ外部の情報に着目し，森ら [11] の Web 検索エンジンを用いたスコア計算法を用いる．森らは，Web 検索エンジンを用いて氏名とキーワードのヒット数を取得し，Jaccard 係数を用いてキーワードの共起の強さを計算している．我々は，森らの手法を人物と複数候補との関連度計算に応用し，関連度の最も高い候補を最終候補とする．

検索氏名 n と複数候補 v の関連度 $J(n,v)$ を以下のように求める．

$$J(n,v) = \frac{|N \cap V|}{|N| + |V| - |N \cap V|} \tag{1}$$

なお，$|N|$ は検索クエリを n として Web 検索エンジンから得られるヒット数，$|V|$ は検索クエリを v として得られるヒット数，$|N \cap V|$ は n と v の AND 検索にて得られるヒット数である．森らは，文脈に応じたキーワード抽出を行うためにコンテクストワードを用いたが，ここでは，職業に関連する語に絞っているため不要である．Web 検索エンジンとしては Google Web APIs [1] を用いる．

最終候補には，職業語，役職，役割の 3 種類がある．最終候補のタイプを判定し，タイプ毎に異なる処理を行い，職業関連情報を作成する．

最終候補が職業語である場合には，それを職業関連情報とする．

最終候補が役職であれば所属と結合する．しかし，「大阪市立大学教授」などのように，役職に所属が含まれている場合もある．そのため，役職に所属が含まれるかどうかを判定し，含まれていなければ所属を結合する．判定には 3.2.2 節 の所属の判定で用いている語を使用する．いずれかの語が最終候補に含まれていれば，所属が既に含まれていると判定し，最終候補を職業関連情報とする．

所属が含まれていないと判定された場合，所属候補クラスタから所属を得る．まず，クラスタ内の候補の出現頻度の合計が最も多いクラスタを選択する．選択されたクラスタ内に所属候補が 1 つであれば，その所属候補

[1] http://www.google.com/apis/

を最終候補と結合し，職業関連情報とする．クラスタ内に所属候補が複数ある場合や最頻度のクラスタが複数ある場合は関連度を計算する．関連度の最も高い所属候補を最終候補と結合し，職業関連情報とする．関連度 $J(n,o,v)$ を，

$$J(n,o,v) = \frac{|N \cap O \cap V|}{|N \cap V| + |O| - |N \cap O \cap V|} \qquad (2)$$

と定義する．n は検索氏名，o は所属候補，v は最終候補（役職）である．$|N \cap V|$ は n と v の AND 検索にて得られるヒット数，$|O|$ は o で得られるヒット数，$|N \cap O \cap V|$ は，n と o と v の AND 検索にて得られるヒット数である．v はコンテクストワードであり，ここでは最終候補（役職）を用いる．

最終候補が役割であれば，著作候補から関連の高い著作を選択して結合する．著作候補の関連度の計算には，著作候補（ISBN 番号）と検索氏名との間の文字数（以下，文字距離）を用いて，文字距離が最も短い ISBN 番号を 1 つ選択する．選択された ISBN 番号から Amazon Web サービス[1] を用いて著作タイトルを取得する．著作タイトルと最終候補（役割）を結合して職業関連情報を作成する．

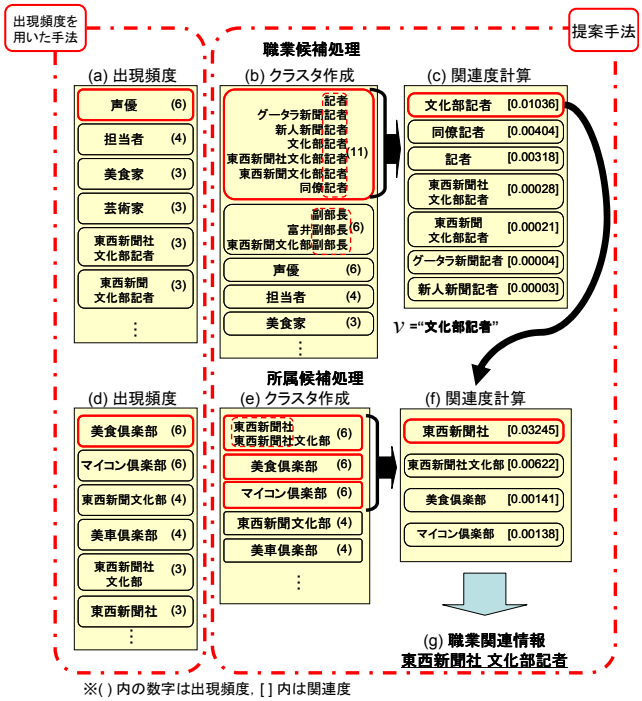


Fig. 5　Examples of Generating Vocation-Related Information using Query “ Shiro Yamaoka ”

3.3.3　処理例

漫画「美味しんぼ」の主人公,「山岡士郎」（職業は新聞記者，東西新聞社勤務）の人物クラスタに含まれる 716 の Web ページを対象とした場合の処理の例について述べる（Fig. 5 参照）．

Fig. 5 (a) は職業関連情報候補抽出処理によって得られた職業語，役職，役割候補であり，この中で最頻度のものは「声優」であるが,「声優」は山岡士郎の職業ではない．職業候補クラスタ作成によって，Fig. 5 (b) の職業候補クラスタが作成され,「記者」を元に作成された 7 つの職業候補を持つクラスタが最頻度となった．このクラスタに含まれる職業候補について関連度計算を行った結果,「文化部記者」の関連度が最も高かった．Fig. 5 (c) ので「文化部記者」を職業候補クラスタから得られた最終候補とする.「文化部記者」は，タイプ判定で役職と判定され，所属を求める必要がある．所属候補の最頻度は「美食倶楽部」であるが，これは山岡士郎の所属する組織ではない（Fig. 5 (d)）．所属候補クラスタを作成したところ，最頻度のクラスタが 3 つ存在し，そのうち 1 つが 2 つの候補を持っているため，関連度計算の対象が 4 候補となる（Fig. 5 (e)）．関連度 $J(n,o,v)$ の v として，最終候補の「文化部記者」を用いた．その結果,「東西新聞社」が最も高い関連度となった（Fig. 5 (f)）．最後に，所属候補「東西新聞社」と役職候補「文化部記者」を結合し，山岡士郎の職業関連情報「東西新聞社 文化部記者」を作成する（Fig. 5 (g)）．

一般的なラベル付け手法を，抽出された名詞から最頻語を選択する手法とすると「声優」が抽出されてしまうが，提案手法では，職業候補のクラスタリング，Web 検索エンジンを用いた関連度計算の効果で「文化部記者」を得ることができている．さらに所属と役職を結合することにより，より人物識別に有用な職業関連情報を作成できている．

4.　実行例

「江川卓」で Google Web APIs を用いて検索し 788 件取得した．この中に何人の同姓同名人物がいるか著者らが判定したところ 3 人いた．どの人物のページか不明なページが 72 あり，この 72 ページを除いた 716 ページを手作業で人物毎に分離した．

分離した Web ページから，提案手法を用いて職業関連情報を抽出した．Table 2 に，各人物の Web ページ数，Web ページを見て人手で抽出した職業関連情報，提案手法で抽出された職業関連情報を示す．

Table 2　Results for Query “ Suguru Egawa ”

	Web ページ数	人手で抽出した職業関連情報	抽出された職業関連情報
江川卓 1	589	野球解説者, キャスター, タレント	野球選手
江川卓 2	126	ロシア文学者, 東京工業大学名誉教授	ロシア文学者
江川卓 3	1	評論家	評論家

[1] http://aws.amazon.com/

江川卓 1 に関しては，提案手法では「野球選手」（職業語,「選手」で判定）と抽出された．これは前職であり，現職は野球解説者，タレントなどである．江川卓 2 の人物（故人）は，ロシア文学者であり，東京工業大学の名誉教授であった．提案手法では「ロシア文学者」（職業語,「学者」で判定）と抽出された．江川卓 3 については，Web ページ中に,「今月末の大手サイト審査会で評論家の江川卓氏を座長とした審査を経て決定される予定」との記述があった．この人物は，江川卓 1 の可能性も考えられるが，Web ページ中に野球に関する記述が全くなかったため別人と判断した．提案手法では,「評論家」（職業語,「家」で判定）が抽出された．

5. 評価実験

提案手法の有効性を確認するために，実験を行った．

実験 1 では，提案手法で得られる職業関連情報が職業に関連する語であると被験者が認知するかどうかを調査した．実験 2 では，提案手法の組み合わせの有効性を確認するため，提案手法と 2 つの比較手法について評価を行った．実験 3 では，提案手法と他の抽出手法との比較評価を行った．

なお，実験での被験者は全て情報工学・情報学を専門とする大学院生である．

Table 3 Dataset

氏名	同姓同名人物数	Web ページ数
竹内郁雄	21	593
和田英一	24	326
野村紀子	27	281
木下和彦	46	243
菱沼聖子	4	293
田中克己	55	330
中村紘子	7	728
上田次郎	6	628
山岡士郎	4	719
新垣紀子	11	339
江川卓	3	716
三浦麻子	22	500

5.1 データセット

先行研究（佐藤ら [2] ）で用いられた 20 の氏名のうち，同姓同名人物が存在する 12 の氏名を採用し，Google Web APIs にて検索を行った．取得した Web ページについて，何人の同姓同名人物が存在するかを調査したところ，延べ 230 人であった．アクセス不能であったページと，どの人物のページか判定できなかったページを取り除き，手作業で人物毎に Web ページを分類した．

作成されたデータセットの詳細を Table 3 に示す．

5.2 実験 1

実験 1 では，提案手法で得られる職業関連情報を，実際に被験者が職業に関連する語であると認知するかどうかを調査した．

5.2.1 方法

全データセット（12 の氏名に含まれる延べ 230 人の人物の Web ページ）から提案手法を用いて職業関連情報を抽出した．抽出できた職業関連情報は 114（49.6%）であった．

114 の職業関連情報を被験者 5 名に提示し，各職業関連情報について，3 段階（2: 職業を表す語であると思う，1: 職業を表す語ではないが職業に関連する語であると思う，0: 全く職業に関連しない語であると思う ）で評価させた．この評価尺度を関連度と呼ぶ．また，2 と 1 を適合とみなし,「提示された語の職業関連情報としての精度」を算出する．

5.2.2 結果と考察

2（職業を表す語）と答えた割合が，62.3%(355/570) ，1（職業を表す語ではないが職業に関連する語）と答えた割合が 25.1%(143/570) ，0（全く職業に関連しない語）と答えた割合が，12.6%(72/570)であった（Fig. 6 参照）．2+1 より，抽出された職業関連情報の 87.4%(498/570) が職業と関連があると評価されたと考える．2 と 1 を適合とみなすと,「提示された語の職業関連情報としての精度」が 87.4%であった．

また，(1) 職業を表す語，(2) 所属と役職を表す語，(3) 著作と役割を表す語別に集計を行った．

(1) の職業を表す語については，2+1 で 83.8%(67/80) が職業に関連があると評価された．一部「イエス・キリスト」や「メーリングリスト」など，職業と全く関連のないものが抽出されたこと，抽出された語が相対的に少ないことより，全体と比べて，全く職業に関連しないと判定された割合が高かった（16.3%）．

(2) の所属と役職を表す語では，2+1 で，88.5%(416/470) と，9 割近くが職業に関連があると評価され，(1) よりも良かった．

(3) の著作と役割を表す語では，2+1 で 75.0%(15/20) であった．(3) で抽出される語は他の 2 つと比べて低かった．2（職業を表す語）として評価される語は存在しなかった．

これらの結果は，本研究で定義した 3 種類の職業関連情報の有効性や，HTML 構造に着目した名詞の抽出手法，簡単なヒューリスティックを用いた職業関連候補判定手法の有効性を示す結果である．特別な職業辞書やルールを用いず，また，機械学習を行わないわりには一定の精度を示しているといえよう．

5.3 実験 2

実験 2 では，提案手法の組み合わせの有効性を確認するために，提案手法と比較手法 1（Web ページ全体を使

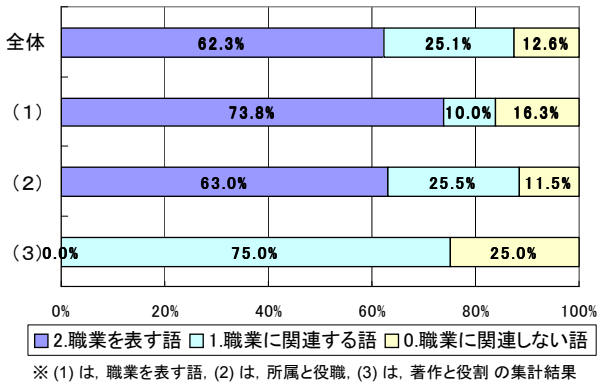


Fig. 6 Results of Experiment 1

用），比較手法 2（出現頻度により最終候補を選択）の 3 手法を用いた比較実験を行った．

5.3.1 比較手法

提案手法と比較する 2 つの手法について述べる．

比較手法 1 は，Web ページ全体を対象とする手法である．Web ページからタグを除去し，名詞（単名詞または複合名詞）を抽出する．名詞抽出後の処理は提案手法と同じである．

比較手法 2 は，出現頻度のみで職業関連情報の候補を選択する手法である．職業関連情報候補抽出処理は提案手法と同じである．職業関連情報の候補について頻度計算を行い，最頻度のものを最終候補とする．最終候補が職業語の場合，そのまま職業関連情報として用いる．役職の場合，所属が含まれるか判定し，含まれていればそのまま職業関連情報とする．含まれていなければ，所属候補から最頻度のものを取得して役職と結合し，職業関連情報を作成する．役割の場合は，著作候補と検索氏名との文字距離を計算し，文字距離の最も短いものから得た著作と役割を結合し職業関連情報とする．

5.3.2 方法

データセットの中から，同姓同名人物数が中程度かつ同程度で，全て大学教員 (研究者) が含まれている，「竹内郁雄」，「和田英一」，「三浦麻子」を選択した．3 つのデータセットについて 3 手法を用いて職業関連情報を抽出した．3 手法の全てで抽出できた人物 32 人を評価の対象とした．

被験者 3 名に対し以下の手順で実験を行った．

手順 1 同姓同名人物毎に分けられた URL リストを閲覧させ，Web ページに含まれる各人物の職業を表す語と職業に関連する語を人手で抽出させた．

手順 2 抽出させた職業を表す語と職業に関連する語を参考に，3 手法で抽出された職業関連情報の関連度を評価させた．

評価尺度は実験 1 と同じである．また，2+1 を適合とみなし，「提示された語の人物に対する職業関連情報としての精度」を算出する．

5.3.3 結果と考察

関連度は提案手法が平均 1.19，比較手法 1 が平均 0.36，比較手法 2 が 平均 1.09 であった．提示された語の人物に対する職業関連情報としての精度は，提案手法で 71.1%(64/90) [1]，比較手法 1 で 24.4%(22/90)，比較手法 2 で 68.9%(62/90)，であった（Table 4 参照）．

比較手法 1 は，他の 2 手法と比べてかなり評価が低かった．提案手法は，3 手法の中で最も良い結果であった．

これらの結果は，提案手法の組み合わせの一定の有効性を示すものであると考える．

提案手法では，Web ページの中から該当人物に関連のある職業の出現しやすい箇所を示し，その有効性を実験により確認できた．比較手法 2 では，職業関連情報候補のランキング手法として出現頻度のみを用いている．提案手法では，候補のクラスタ作成を行った後に，出現頻度と Web 検索エンジンのランキングを用いている．

人物クラスタの Web ページ数が少ない人物では，提案手法と比較手法 2 で全く同じ職業関連情報が作成されることが多かったが，Web ページ数が多い人物では，提案手法の方が比較手法 2 よりも良い職業関連情報が作成されることが多かった．たとえば，山岡士郎（東西新聞社文化部勤務の新聞記者）の処理例にあげたとおり，提案手法では「東西新聞社文化部記者」，比較手法 2 では「声優」と出力された．ここで，Web ページ数が多い人物の場合（11 ページ以上）と少ない人物の場合（10 ページ以下）に分けて関連度の比較を行った．その結果，Web ページ数が多い人物の場合，提案手法で関連度 1.83，比較手法 1 で 1.11，比較手法 2 で 1.44，少ない人物の場合，提案手法で 0.95，比較手法 1 で 0.15，比較手法 2 で 0.92 であった．このように，Web ページの数によらず提案手法が最も良いが，Web ページ数が多い人物の場合の方が差が大きかった．Web ページ数が多い人物の場合，出現頻度のみによるランキング手法よりも，クラスタ作成と Web 検索エンジンを用いた関連度計算手法を用いる方が有用な職業関連情報を抽出できると考える．

Table 4 Results of Experiment 2

	提案手法	比較手法 1	比較手法 2
関連度	1.19	0.36	1.09
精度	71.1%	24.4%	68.9%

5.4 実験 3

提案手法と，先行研究である Wan ら [6] の肩書抽出手法と，一般的に人物の職業に関する情報が含まれる可能性が高いと考えられるプロフィールページから抽出する手法（以下，プロフィールページ抽出手法）との比較実

[1] 対象人物のうち 2 人が評価中に Web ページの閲覧ができなかったため，集計から除外した．

験を行った．

5.4.1 比較手法

(1) 肩書抽出手法

Wan らは，肩書を 13 種類のヒューリスティックを用いて抽出し，出現頻度の最も多いものをラベルとしている．しかし，論文中では 13 種類のヒューリスティックのうち 1 種類しか紹介されていない．たとえば「David Lee is a painter」という文があった場合，「painter」を肩書とするものである．本実験ではこのヒューリスティックを肩書抽出手法とし，Fig. 7 に示す．

Step.1 Web ページに含まれる文章を 1 文毎に切り出す．

Step.2 切り出された文に対し，CaboCha[1] を実行し文節毎に切り分ける．

Step.3 切り分けた文節に，ChaSen を実行する．

Step.4 検索氏名または苗字と「助詞-係助詞 」の「は」が隣接する文節と，名詞と「助動詞 特殊」と識別された語が隣接する文節の両方を含む文を抽出する．

Step.5 抽出された文から「助動詞 特殊」と識別された語の前に出現する名詞を職業関連情報候補とする．

Step.6 最頻度の職業関連情報候補を職業関連情報とする．

Fig. 7 Comparative Method: Extracting Titles

(2) プロフィールページ抽出手法

人物に関連するプロフィール・経歴が紹介されているページから職業関連情報を抽出する．ページの大部分で人物に関するプロフィールや紹介が書かれているページをプロフィールページと定義した．

まず，大学院生 3 名にデータセットに含まれる全ての Web ページを閲覧させ，その中から各人物のプロフィールページを抽出させた．抽出したプロフィールページから，各人物に最適な職業を表す語，または職業に関連する語を 1 つ抽出させた．

5.4.2 方法

データセット全てを用いた．

被験者 3 名を対象に，4 セットずつ割り当て関連度を評価させた．評価手順，評価尺度は，実験 2 と同じである．提示された語の人物に対する職業関連情報としての精度も同様に算出した．

次に，再現率を以下のとおり求めた[2] ．

$$\text{再現率} = \frac{\text{適合と判断された職業関連情報を抽出できた人物数}}{\text{適合と判断される職業関連情報を人手で抽出できる人物数}}$$

[1] http://www.tahoo.org/~taku/software/cabocha/

[2] ただし，分母については，Web ページから適合する職業関連情報を抽出できるかどうかの判定は非常に困難であるため，本研究での職業関連情報の定義に基づき著者らが行った．

Table 5 Results of Experiment 3

	提案手法	肩書抽出手法	プロフィールページ抽出手法
関連度	1.41	1.17	1.76
精度	84.7%	66.7%	90.5%
再現率	66.7%	2.8%	13.5%

5.4.3 結果と考察

関連度は提案手法で平均 1.41，肩書抽出手法で平均 1.17，プロフィールページ抽出手法で平均 1.76 であった．提示された語の人物に対する職業関連情報としての精度は提案手法で 84.7%(94/111)，肩書抽出手法で 66.7%(4/6)，プロフィールページ抽出手法で 90.5%(19/21) であった．関連度と精度については，プロフィールページ抽出手法が最も良かった．自動の手法の中では提案手法の結果が良かった．関連度と精度において，プロフィールページ抽出手法が最も良い理由の一つは，人手で行っているからでもある．自動で行う場合には若干精度は低下するであろう．また，プロフィールページはどの人物にも存在するものではない．

再現率は，提案手法で 66.7%(94/141)，肩書抽出手法で 2.8%(4/141)，プロフィールページ抽出手法で 13.5%(19/141) であり，提案手法が特に良かった（Table 5 参照）．

以上の結果は，精度と再現率を総合的に見た提案手法の有効性を示していると考える．

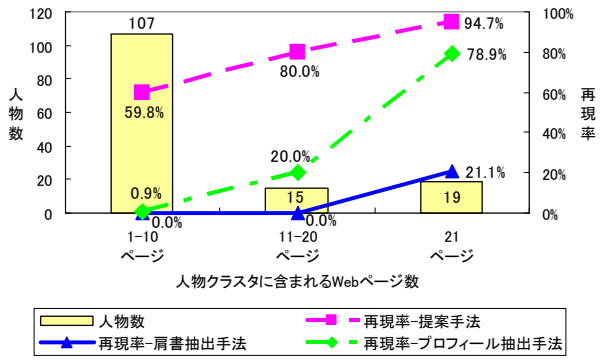


Fig. 8 Number of People and Recall by Number of Web Pages in a Person Cluster

Web ページ数別の人物数と再現率を集計した（Fig. 8 参照）ところ，Web ページ数が 10 ページ以下の人物が，全体で 75.9%(107/141) を占めていた．10 ページ以下の人物に限って集計したところ，再現率は提案手法では 59.8%(64/107) であり，肩書抽出手法では 0%(0/107)，プロフィール抽出手法では 0.9%(1/107) であった．Web ページ数が少ない人物の場合に，提案手法が特に優れていると考える．しかし，提案手法でも Web ページ数が多い人物の場合と比べて再現率が低く，改善を図る必要がある．

そこで，Web ページ数が 10 ページ以下の人物のうち，再現率の計算において適合と判断される職業関連情報を抽出されなかった 43 人物について理由を分析した．何も抽出されなかった人物が 72.1%(31/43),「0（全く職業に関連しない語)」（不適合）と判定された人物が 27.9%(12/43) であった．前者の 31 人物について Web ページを詳細に見たところ，61.3%(19 人物) は table 要素を用いた表形式に氏名が羅列されている（例えば会員名簿など）ものであり，表のヘッダーやキャプション部分，ページのタイトルなど，名詞抽出の対象外に職業関連情報が記述されていた．不適合と判定された後者の 12 人物については，主として他人の職業関連情報が抽出されたものであった．一方，Web ページ数が 11 ページ以上の人物については，1 人物で何も抽出されず，3 人物で不適合判定であった．抽出されなかった 1 人物は，名詞抽出の対象外に職業関連情報が記述されており，不適合の 3 人物は 10 ページ以下と同様の理由であった．

すなわち，Web ページ数が少ない人物の再現率を向上させるためには，職業関連情報候補抽出過程における，ヒューリスティック追加などのアルゴリズム改善が必要である．ただし，Web ページ数が多い人物の場合にはほとんど問題が発生していないため，提案手法による簡単な抽出手法は一定の有効性を示していると考える．

6. 関連研究と議論

6.1 関連研究

本研究の目的は Web 上の同姓同名人物の識別を容易にすることである．代表的な先行研究である，Wan ら [6] の研究と比較する．Wan らは，人名検索で得られた英語の Web ページを同姓同名人物毎に分離し，各人物に対しラベル付与を行っている．Web ページから簡単なヒューリスティックを用いて肩書を抽出し，出現頻度の最も多いものをラベルとして用いている．抽出される肩書はおおむね 1 つか 2 つの英単語である．本研究では以下の点が異なる．日本語のページを対象としている．肩書ではなく 3 種類の職業関連情報をラベル付けしている．Web ページに含まれる語を組み合わせることにより職業関連情報を作成している．出現頻度，同義クラスタの作成，Web 検索エンジンを用いた関連度計算を行っている．先行研究ではたとえば「professor（教授）」のような 1 語の肩書が表示されるが，本研究では，該当人物の Web 上での言及に応じて「大阪市立大学教授」「工学研究者」などが表示される．

Web 上の同姓同名人物に関する研究は，同姓同名人物の分離がほとんどである．たとえば，佐藤ら [2] ，Bekkerman ら [3] ，白砂ら [4] ，木村ら [5] の研究がある．佐藤ら，Bekkerman らの研究は，人間関係を用いた同姓同名人物の分離を行っている．白砂らは職業を含むプロファイル情報に着目し，木村らは Web 検索エンジンのスニペットに着目した手法を提案している．本研究は同姓同名人物の分離後の人物クラスタの識別を対象としている．なお，白砂らは分離（クラスタリング）に用いる特徴語の一つとして，辞書を用いて人物の職業を抽出しているが，本研究とは抽出手法が異なる．

Web ページから人物に関する情報を抽出する研究には，キーワードを抽出する研究 [11] ，呼称や別名を抽出する研究 [12,13] ，経歴を抽出する研究 [14] などがある. [11] は，語の共起を利用してキーワード抽出を行っている. [12,13] は，人物に関する呼称や別名がよく出現すると考えられるパターンを利用し抽出を行っている. [13] に関しては，パターンの自動生成も試みている. [14] は，経歴情報として日付表現とその日付に付随する人物情報を収集している．本研究では職業関連情報の抽出を行っている．職業候補クラスタと所属候補クラスタから候補を選び出す段階で，森らが提案した手法 [11] を応用している．

本研究は，Web 検索の分類（クラスタリング）結果のラベル付けとも関連する．一般的なラベル付け手法は，Web ページに含まれる特徴語を抽出して，出現頻度に基づくランキングを行うものである．TSUBAKI[15] では，特徴語の抽出時，複合名詞のクラスタ作成を行っている点が類似しているが，Web 検索エンジンを用いたランキングを行っていない点が異なる．

6.2 議論

本研究では，Web 上の同姓同名人物の識別を容易にするために，日本語の Web ページから職業関連情報を抽出する手法を提案した．

本研究の特徴は，厳密に定義された職業を出現頻度に応じて付与するのではなく, 人物の識別に有用と思われる職業関連情報を文脈に応じて付与するという着眼点にある．たとえば, 研究者であり教員でもある大学教員の場合, 職業である「研究者」や「(大学）教員」ではなく, 該当人物が Web 上で出現する文脈に応じて「大阪市立大学教授」「情報工学研究者」などを出力できる．特別な職業辞書が不要であることも大きな利点である．

評価実験の結果, 以下の知見を得た．

(1) 提案手法で抽出した語の職業関連情報としての精度は 87%であった．提案手法で抽出した語の人物に対する職業関連情報としての精度は 71%であった．これらは, 提案手法の一定の有効性を示している．

(2) 提案手法は全自動の比較手法の中で, 関連度, 精度において最も良かった．提案手法は手動の比較手法（プロフィールページからの抽出手法）には関連度, 精度においてやや劣るが, 再現率は 66.7%であり, 手動の比較手法 (13.5%) より大幅に良かった．これらは, 提案手法の全自動の手法としての有効性を示している．

(3) Web ページ数が多い人物の場合, 提案手法では, 頻

度だけを用いる手法と比べて良い職業関連情報を抽出できる．Web ページ数が少ない人物の場合，比較手法ではほとんど職業関連情報を抽出できないが，提案手法ではある程度職業関連情報を抽出できる．これらは，提案手法の長所を示している．

以下に今後の課題をまとめる．

まず，職業関連情報の候補抽出の精度改善が必要である．不要語辞書や，機械学習による手法を検討したい．

次に，Web ページ数が少ない人物の場合に再現率の向上が必要である．職業関連情報が抽出されなかった場合で Web ページ数が少ない人物の場合には，ヒューリスティックを追加した抽出アルゴリズムを実行するなどの改善を試みたい．

また，提案手法では，過去の職業が抽出されることがある．実験の結果，関連度は高い．前職で有名な場合（例：江川卓１）や，昇任や転職したばかりである場合は，妥当な結果であるともいえる．しかし，多くの場合において，前職よりも現職を期待されるのではないだろうか．技術的には，[14] のように Web ページ上の時系列情報を考慮に入れることにより解消できるのではないかと考えるが，今後の課題である．また，どちらを提示すべきか「指標」に関しても，今後の検討課題である．

本研究ではラベル付けとして職業関連情報に焦点をあてたが，人物クラスタの選択支援のためには，職業関連情報が抽出されない場合の処理を検討する必要がある．一つは，先行研究と同様に，代替情報として人物クラスタの特徴語を提示することである．他には，特徴語，出生年，地名などのプロフィール情報を提示する方法が考えられる．

7. おわりに

Web 上の同姓同名人物の識別を容易にするため，人物毎に分離された Web ページから職業関連情報を抽出する手法を提案した．本研究における職業関連情報とは，(1) 職業を表す語，(2) 所属と役職を表す語，(3) 著作と役割を表す語，である．HTML 構造に着目した名詞の抽出，簡単なヒューリスティックを用いた職業関連候補判定，出現頻度と同義クラスタ作成と Web 検索エンジンを用いたランキングに基づく職業関連情報作成手法を提案した．

本研究の特徴は，厳密に定義された職業を出現頻度に応じて付与するのではなく，文脈に応じて人物の識別に有用と思われる職業関連情報を付与するという着眼点にある．特別な職業辞書が不要であることも大きな利点である．

実験の結果，抽出された語の職業関連情報としての精度が 87% ，抽出された語の人物に対する職業関連情報としての精度が 71% であり，一定の有効性を確認した．他の手法と比べての有効性も確認した．

最後に，提案手法は，Web からの人物プロフィール情報の自動作成に応用できると考える．

参 考 文 献

[1] R. V. Guha and A. Garg: Disambiguating People in Search, Standord University (2004)

[2] 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, pp. 26–36 (2005)

[3] R. Bekkerman and A. McCallum: Disambiguating Web Appearances of People in a Social Network, Proc. of the 14th World Wide Web Conference(WWW2005) (2005)

[4] 白砂 健一, 小山 聡, 田島 敬史, 田中 克己: Web の構造情報とプロファイル抽出を用いたオブジェクト識別, 第 17 回データ工学ワークショップ論文集 (DEWS2006), 2C-i7 (2006)

[5] 木村 塁, 戸田 浩之, 田中 克己: 検索結果スニペットのクラスタリングによる同姓同名人物の特定, 第 17 回データ工学ワークショップ論文集 (DEWS2006), 2C-i11 (2006)

[6] X. Wan and J. Gao, M. Li and B. Ding: Person Resolution in Person Search Results: WebHawk, Proc. of the 14th ACM international conference on Information and knowledge management, pp. 163–170 (2005)

[7] 新村 出 (編): 電子版 広辞苑 第五版, 岩波書店 (1998)

[8] 労働省 職業安定局 (編): 平成 11 年改定 労働省編 職業分類 －職業分類表－, 雇用情報センター (2000)

[9] 上田 洋, 村上 晴美: Web 上の同姓同名人物を分離して人物属性情報を表示するシステム, 第 21 回人工知能学会全国大会, 3G8-1 (2007)

[10] 渡辺 一郎, 桝井 文人, 福本 淳一: 固有表現抽出ツール NExT の精緻化とユーザビリティの向上, 第 10 回言語処理学会年次大会発表論文集, pp. 413–415 (2004)

[11] 森 純一郎, 松尾 豊, 石塚 満: Web からの人物に関するキーワード抽出, 人工知能学会論文誌, Vol. 20, No. 5, pp. 337–345 (2005)

[12] 外間 智子, 北川 博之: Web データを用いた人物の呼称抽出, 日本データベース学会論文誌 (DBSJ Letters), Vol. 5, No. 2, pp. 49–52 (2006)

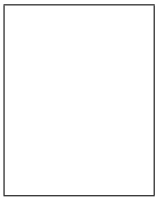
[13] 本間 大輝, Danushka Bollegala, 松尾 豊, 石塚 満: Web を用いた人物の別名抽出, NLP 若手の会第 2 回シンポジウム (2007)

[14] 木村 塁, 小山 聡, 田中 克己: Web からの人物事典生成のための経歴情報の自動収集, 日本データベース学会論文誌 (DBSJ Letters), Vol. 5, No. 2, pp. 29–32 (2006)

[15] 馬場 康夫, 新里 圭司, 黒橋 禎夫: 検索エンジン基盤 TSUBAKI を用いた大規模ウェブ情報クラスタリングシステム, 情報処理学会研究会報告, 2008-FI-89/2008-NL-183, pp. 67–74 (2008)

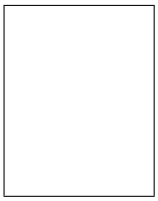
著 者 略 歴

上田 洋 (学生会員)



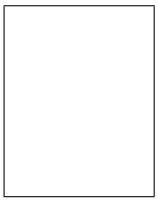
2005年3月大阪市立大学大学院創造都市研究科都市情報学専攻修士課程修了. 2006年4月大阪市立大学大学院工学研究科電子情報系専攻後期博士課程入学, 現在に至る. 情報処理学会, 人工知能学会, 日本図書館情報学会の会員.

村上 晴美



1986 年京都大学文学部哲学科心理学専攻卒業, 富士通株式会社入社. 1994 年英国 UMIST 計算機学科修士課程修了. 1998 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了, 大阪市立大学学術情報総合センター講師. 2001 年同助教授. 2003 年大阪市立大学大学院創造都市研究科助教授. 2007 年同教授となり現在に至る. M.Sc., 博士 (工学). テキスト・データからの人物の理解に関する研究に従事. 情報処理学会, 人工知能学会, 日本認知科学会, 日本図書館情報学会などの会員.

辰巳 昭治



1970 年 3 月大阪大学工学部通信工学科卒業. 1972 年 3 月同大学大学院工学研究科通信工学専攻修士課程修了. 1972 年 4 月川崎重工業（株）入社. 1978 年 3 月大阪大学大学院工学研究科通信工学専攻博士課程修了. 工学博士. 豊橋技術科学大学を経て, 現在, 大阪市立大学大学院工学研究科電子情報系専攻教授. パターン認識と学習に関する研究、並列計算モデルの研究に従事. 電子情報通信学会, 情報処理学会, 人工知能学会, IEEE,ACM などの会員.