特集論文

Web 上の人物理解のための履歴書作成

Creating Curriculum Vitae for Understanding People on the Web

上田 洋 Hiroshi Ueda
大阪市立大学大学院工学研究科
Graduate School of Engineering, Osaka City University
d06tb001@ex.media.osaka-cu.ac.jp, http://kuroyuri.media.osaka-cu.ac.jp/~ueda/

村上 晴美 Harumi Murakami
大阪市立大学大学院創造都市研究科
Graduate School for Creative Cites, Osaka City University
harumi@media.osaka-cu.ac.jp, http://www.media.osaka-cu.ac.jp/~harumi/

辰巳 昭治 Shoji Tatsumi
大阪市立大学大学院工学研究科
Graduate School of Engineering, Osaka City University
tatsumi@info.eng.osaka-cu.ac.jp, http://www.kdel.info.eng.osaka-cu.ac.jp/~tatsumi/

keywords: curriculum vitae, web people search, information extraction, classification, clustering

Summary

When users find information about people from the results of Web people searches, they often need to browse many obtained Web pages and check much unnecessary information. This task is time-consuming and complicates the understanding of the designated people. We investigate a method that integrates the useful information obtained from Web pages and displays them to understand people. We focus on curriculum vitae, which are widely used for understanding people. We propose a method that extracts event sentences from Web pages and displays them like a curriculum vita. The event sentence includes both time and events related to a person. Our method is based on the following: (1) extracting event sentences using heuristics and filtering them, (2) judging whether event sentences are related to a designated person by mainly using the patterns of HTML tags, (3) classifying these sentences to categories by SVM, and (4) clustering event sentences including both identical times and events. Experimental results revealed the usefulness of our proposed method.

1. はじめに

近年，インターネットの普及により，Web 上から様々な情報が得られるようになり, Web 検索エンジンを用いることが一般的となった．[Guha 04] 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, Web での人名を用いた検索（以下，Web 人名検索）のニーズは高いといえる．

Web 人名検索の目的は 2 つに大別されると考える．1 つは，(1) よく知らない人物についてどのような人物かを知るため，もう 1 つは，(2) 既にある程度知っている人物についてより詳しい情報を知るためである．(1) の場合，ある程度の数の Web ページを閲覧しなければ検索対象の人物の全容がつかめない可能性が高いが, 多数の Web ページの閲覧は利用者の負担となる．また, 閲覧する Web ページの中には, 人物が一時的に仕事で立ち寄った事実の記録など, 該当人物の理解にあまり大きく役立たない情報も含まれる．よく知らない人物については，Web 人名検索結果から人物の理解に役立つ情報を抽出，統合して提示することにより，どのような人物かを知るための利用者の負担を軽減できると考える．また，(2) の場合についても，提示された情報の中に利用者の求める情報があれば，Web ページ内を探すことなくすばやく情報を入手できる．

本研究では, Web 人名検索結果から人物の理解に役立つ情報を統合して提示する手法を検討する．何を人物の理解に役立つ情報とするか，また，その統合や提示方法は自明ではない．我々は，人物の理解に役立つ情報の統合・提示方法として, 多くの人になじみのある履歴書に着目した．履歴書に記載される履歴を人物の理解に役立つ情報と考える．履歴には，氏名，生年月日，学歴，職歴，賞罰などがあり，多くは「2005 年 3 月 31 日大阪市立大学卒業」のような，時間と人物に関する出来事の両方を含む文である．そこで，本研究では,「時間と，人物に関する出来事の両方を含む文」を履歴文と呼び，Web ページから抽出する．履歴書には学歴，職歴などのカテゴリがある．本研究でも，履歴文を学歴や職歴などのカテゴリ毎に分類して履歴書の形式で提示する．

本研究では，Web 人名検索結果の Web ページから, 人物の理解に役立つ情報を抽出し, 履歴書の形式で提示する手法を提案する．履歴文を抽出し，戸籍（生，没年月日等の主に戸籍に関係するもの），学歴，経歴，受賞歴

のカテゴリ毎に分類して提示する．

本論文の構成は以下の通りである．2 章で提案手法を説明する．3 章で提案手法の実行例を示す．4 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．5 章で関連研究と比較して議論し，6 章でまとめる．

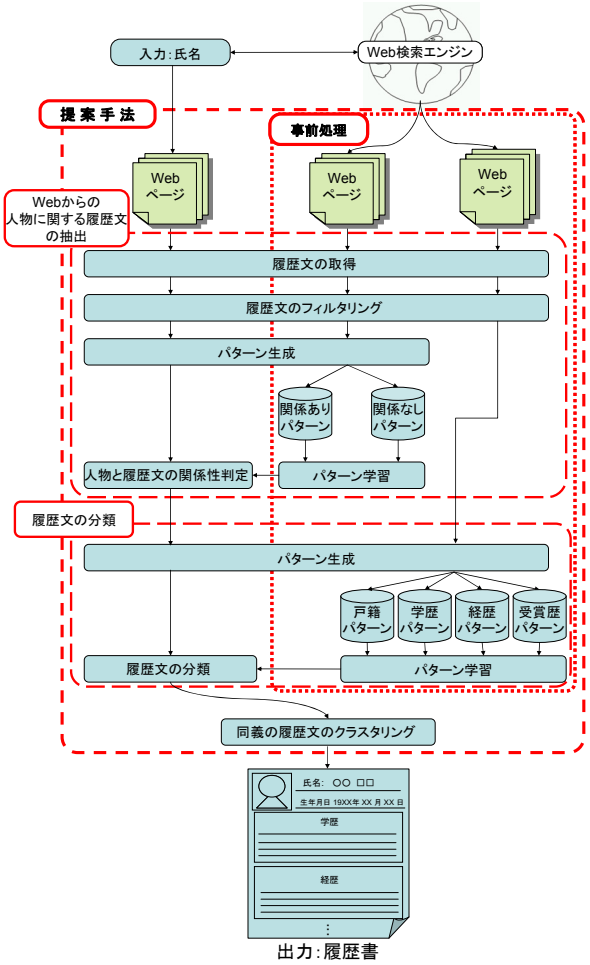


図 1 提案手法概要

2. 提 案 手 法

本研究では，Web ページから履歴文を抽出し，履歴書の形式で提示する手法を提案する．

2·1 概 要

Web 人名検索にて得られた HTML で記述された Web ページから履歴文を抽出し，カテゴリ毎に分類し履歴書の形式で提示する．提案手法は，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．提案手法の概要を図 1 に示す．

Web からの人物に関する履歴文の抽出では，まず，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．

履歴文の分類では，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（戸籍，学歴，経歴，受賞歴）に分類する．

同義の履歴文のクラスタリングでは，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．

2·2 Web からの人物に関する履歴文の抽出

Web からの人物に関する履歴文の抽出では，Web ページからヒューリスティックにより履歴文を取得し，不要語と不要パターンを用いて履歴文のフィルタリングを行う．フィルタリング処理した履歴文について，人物と関係があるかを主に HTML のタグの出現パターンを用いて SVM により判定する．

§1 履歴文の取得

HTML で記述された Web ページから履歴文を取得する．本研究では，西暦か年号の形式で記述された年（例えば，2008 年，’08，平成 20 年など）が出現する文を履歴文とする．ヒューリスティックを用いて，Web ページから文を切り出し，得られた文に年を含むかどうかを判定する．

文の切り出しに用いるヒューリスティックとして，以下の 3 つを選定した．

1. 句点「。」で終わる 1 文
2. tr, li, h1〜h5, title, p, div タグで囲まれ，句点を含まない文字列
3. br タグで終わり，句点を含まない文字列

ただし，1. と 3. の開始位置は，直前に出現する句点か br タグ，または 2. であげたタグの直後である．

得られた文に年を含むかどうかを判定するためのヒューリスティックとして，以下の 3 つを選定した．

1. 数字が 4 つ連続する文字列（例：2008）
2. 「 ’」に続いて数字が 2 つ連続する文字列（例：’08）
3. 和暦[*1]に続き数字が 1 つまたは 2 つ現れ，続いて「年」がつく文字列（例：昭和 60 年，平成 20 年）

§2 履歴文のフィルタリング

得られた履歴文から，該当人物に関係のない履歴文や，履歴書に記載不要な履歴文を除去するために，不要語と不要パターンを用いて履歴文のフィルタリングを行う．

*1 明治以降を対象とした．

不要語として 21 種類を選定した．例えば「許諾」「転載」「投稿」「レビュー」「copyright」などである．これらは，商用サイトや掲示板，ブログなどで機械的に生成される定型文に頻出する．

不要パターンとして 20 種類を選定した．例えば「2008 年 12 月 31 日 12 時 00 分」「2008-12-31 12:00:00」などのように, 詳細な時間を表すものである. 詳細な時間の多くは，Web ページやブログの更新時間，コメント，掲示板の投稿の時間などである．

フィルタリングにより除外される履歴文のほとんどは，人物と関係のないものであり，人物に関係がある場合でも，履歴書に記載する内容が含まれる可能性は低いと考える．

§3 人物と履歴文の関係性判定

本研究では，検索対象の人物の氏名を含む Web ページを使用するが，その Web ページから得られる情報全てが検索対象の人物と関係があるとは限らない．そこで，検索対象の人物と関係のある履歴文かどうかを判定する．

人物と履歴文の関係性判定は，文書内のオブジェクト同士の関連性を求める研究と類似し，従来手法としてプレーンテキストを対象とする手法や，DOM などの木構造を用いる手法などがある．

プレーンテキストを対象とする手法は，自然言語処理を用いて関連性を求めることが一般的だが，HTML では文にならない記述が多くあり, 自然言語処理だけでは不十分である．

DOM などの木構造を用いる手法は，HTML のタグの階層構造を利用するため，特に構造的に記述されたページで有効性を発揮する．しかし，Web 人名検索で得られる HTML で記述された Web ページは雑多であり，構造的に記述されているとは限らない．タグの閉じ忘れ等，記述に間違いのあるページもある．木構造の取得には HTML 全体の論理構造の解析が必要となるため，HTML のタグに記述ミスがある場合，有効に機能しない．例えば，table タグが閉じ忘れられている（対応する </table> がない）場合，タグの終点がどこかが不明となり，タグ中の木構造が得られない．table タグが何重にも入れ子になっている場合などには，Web ページ全体の木構造の取得に失敗する可能性もある．

本研究では，人物の氏名または姓と履歴文の間の各タグの出現パターンを主に用いて，SVM により判定する手法を提案する．提案手法であれば，DOM などの木構造を用いる手法では有効性が限定的であった，あまり構造的に記述されていない場合や，HTML の記述ミスがある場合にも対応できる．例えば，上述の table タグが閉じ忘れられた場合でも判定可能である．

人物と履歴文の関係性判定で用いる特徴を表 1 に示す．表 1 の 1 から 4 と 20，21，23，24，26，27，29，30 は，プレーンテキストで書かれた部分を対象とする特徴である．表 1 の 5 から 13 では，タイトルタグやヘディングタグなど Web ページ内の情報に広く影響を与えると考えられる部分について特徴を取得している．表 1 の 14 から 19 は，表 (table) の中にあるものについてのみ対象とする．表 1 の 22，25，28，31 では，人物の氏名または姓と履歴文の間の各タグの出現数を特徴とする．対象とするタグを表 2 に示す．これらのタグは，改行の意味を含み，タグの前後で意味の違いが発生する場合が多いと考える．

表 1 人物と履歴文の関係性判定にて学習のパターンに使用する特徴

1. 履歴文に氏名を含む
2. 履歴文に姓を含む
3. 履歴文に含まれる名詞の数
4. 履歴文に含まれる対象人物以外の人名の数
5. title タグ内に氏名を含む
6. title タグ内に姓を含む
7. title タグ内の名詞の数
8. 最初に出現する h1 タグ内に氏名を含む
9. 最初に出現する h1 タグ内に姓を含む
10. 最初に出現する h1 タグ内の名詞の数
11. 履歴文の最も近くに出現する h1〜5 タグ内に氏名を含む
12. 履歴文の最も近くに出現する h1〜5 タグ内に姓を含む
13. 履歴文の最も近くに出現する h1〜5 タグ内の名詞の数
14. table タグ内の最初に出現する tr タグ内に氏名を含む*3
15. table タグ内の最初に出現する tr タグ内に姓を含む*3
16. table タグ内の最初に出現する tr タグ内の名詞の数*3
17. 履歴文が出現する tr タグ内の最初の td タグ内に氏名を含む*3
18. 履歴文が出現する tr タグ内の最初の td タグ内に姓を含む*3
19. 履歴文が出現する tr タグ内の最初の td タグ内の名詞の数*3
20. 履歴文とその前方の最も近くに出現する氏名の間の名詞の数
21. 履歴文とその前方の最も近くに出現する氏名の間の対象人物以外の人名の数
22. 履歴文とその前方の最も近くに出現する氏名の間の各タグの出現数
23. 履歴文とその前方の最も近くに出現する姓の間の名詞の数
24. 履歴文とその前方の最も近くに出現する姓の間の対象人物以外の人名の数
25. 履歴文とその前方の最も近くに出現する姓の間の各タグの出現数
26. 履歴文とその後方の最も近くに出現する氏名の間の名詞の数
27. 履歴文とその後方の最も近くに出現する氏名の間の対象人物以外の人名の数
28. 履歴文とその後方の最も近くに出現する氏名の間の各タグの出現数
29. 履歴文とその後方の最も近くに出現する姓の間の名詞の数
30. 履歴文とその後方の最も近くに出現する姓の間の対象人物以外の人名の数
31. 履歴文とその後方の最も近くに出現する姓の間の各タグの出現数

注：氏名と姓は対象人物のもの

事前処理として，教師データを作成して SVM により学習させる．まず，政治家，研究者で構成される 5 人の氏名*4 をクエリに Google Web APIs を用いて検索し，各 200 件，計 1000 件の Web ページを取得した．得られた

*3 履歴文が table タグ内にある場合．

*4 人物の職種には偏りがあるが，Web ページの記述パターンに関しては，職種による偏りは小さいと考える．同姓同名の別人の Web ページが検索される恐れがある場合は，該当人物の所属を用いて絞り込みを行っている．

Web ページに対し 2.2 節の§1，2 の処理を行い，7211 の履歴文を取得した．7211 の履歴文に対し，人手により人物と関係があるかどうかを判定した[*5]．その結果，人物と関係のある履歴文（正解データ）2266，人物と関係のない履歴文（不正解データ）4474 に分離し，教師データとした．教師データから表 1 の特徴を抽出してパターンを生成し，SVM [*6] にて学習させた．学習の結果得られたデータを元に，未判定の履歴文のパターンを評価し，人物と関係があるかどうかを判定する．関係があると判定された履歴文のみを以下の処理で用いる．

表 2 解析対象タグ

br	ul
table	ol
tr	li
td	hr
dd	div
p	title
h1-h5	

2·3 履歴文の分類

人物と関係があると判定された履歴文を対象に，カテゴリ毎に分類する．分類には，SVM を用いる．基本的に SVM は二値分類器であるため，多値分類器へ拡張する必要がある．本手法では，多値分類器への拡張手法として one-versus-rest 法を用いる．

分類するカテゴリは，戸籍，学歴，経歴，受賞歴である．カテゴリは，履歴書を参考に選定した．戸籍は，人物の生・没年月日，結婚，国籍変更など戸籍の変更に関する情報を分類する．一般的な履歴書には，戸籍に対応する記述欄はないが，生年月日等，履歴書に個別に記載する情報（以下，個別情報）を抽出するために分類する．個別情報の抽出手法については 2.4 節，具体的な抽出例は 3 章にて述べる．経歴は主に職歴を分類する．本研究では，人物の企業や団体への就職，加入，参加を経歴と定義した．受賞歴には，履歴書の賞罰に当たる情報を分類する．人物の「罰」に相当する情報は記述されることが少ないと考え,「賞」のみを対象に，受賞歴として分類することとした．

事前処理として，戸籍，学歴，経歴，受賞歴の 4 つについて教師データを作成する．まず，これらのカテゴリに当てはまる履歴文を多く含むと考えられる Web ページを取得する．Web ページの取得には，Google Web APIs を用いた．検索に使用したクエリと取得件数を表 3 に示す．使用したクエリの検索結果の大部分は，プロフィールページや人物の履歴を含むページであった．

得られた 500 件の Web ページについて，2.2 節の§1，2 の処理を行い，14974 の履歴文を取得した．14974 の履歴文について，4 つのカテゴリの定義に当てはまるかどうかを人手で判定し分類した．いずれのカテゴリにも当てはまらなかった履歴文は，4 つのカテゴリの学習の際の不正解データとして用いた．全ての履歴文について，形態素解析を行い，TF・IDF を計算する．IDF の文書数は，履歴文の数とする．得られた各形態素の TF・IDF と形態素数を用いて履歴文のパターンを生成する．生成したパターンを用いて，カテゴリ毎に非線形 SVM で学習させた．

表 3 使用クエリと取得件数

クエリ	件数
経歴 site:ja.wikipedia.org	200 件
経歴 inurl:profile	200 件
出身学校 site:read.jst.go.jp	100 件

事前処理にて得られた 4 つのカテゴリの学習の結果を用いて，履歴文を分類する．

2·4 同義の履歴文のクラスタリング

4 つのカテゴリに分類された履歴文の中には，同義の履歴文が存在する可能性がある．重複した意味を持つ履歴文の存在は，無駄な閲覧時間を増やすことになる．そこで，同義の履歴文のクラスタリングを行う．

一般的な類似度のみを用いるクラスタリング手法を使用する場合，時間は異なるが記述された内容が類似した履歴文が同じクラスタになる恐れがある．例えば,「2003 年　大阪市立大学大学院　修士課程　入学」,「2005 年　大阪市立大学大学院　修士課程　修了」の 2 つの履歴文がクラスタリングの対象となったとする．この 2 つの文は，大部分が同じ文字列「大阪市立大学大学院　修士課程」で構成されており，一般的なクラスタリング手法では非常に類似した文として認識され，同じクラスタとしてまとめられる可能性が高い．しかし，これらは「2003 年」「2005 年」と異なる時間に起った出来事であるため，別のクラスタとして分けられなければならない．本研究では,「2003 年　大阪市立大学大学院　修士課程　入学」,「2005 年　大阪市立大学大学院　修士課程　修了」のような比較的短い文を扱うことを想定しており，一般的なクラスタリング手法ではこのような事例が多々発生する．

本研究では，記述された時間を考慮した同義の履歴文のクラスタリングを提案する．提案手法を図 2 に示す．なお，提案手法内の単一パス法では，重み付けに TF・IDF，類似度に余弦を用いる．提案手法であれば，前述の 2 つの履歴文を分離可能である．

同義の履歴文のクラスタリングの副次的な利点として，個別情報の抽出が容易になるという点がある．特に，人生で一度しかない出来事の場合，簡単なヒューリスティックでも高い精度で抽出可能である．例えば，生年月日の場合，時間を考慮したクラスタリングにより得られたクラスタ内は基本的に同じ時間であるため,「生年月日」や「生まれ」のような生年月日に関係する言葉を最も多く含むクラスタの時間を抽出すればよい．生年月日や没年月

*5 人物と関係があるか判定不能のものは除外した．
*6 予備実験の結果，相対的に評価が良かった多項式カーネルを用いた非線形 SVM を使用することとした．

Step.1 年, 月, 日を含む履歴文（以下，年月日データ）を取得し，各年月日データのクラスタを作成する．

Step.2 同年月日のクラスタを結合する．

Step.3 年, 月を含む履歴文（以下，年月データ）のうち，クラスタと同年月の年月データを取得し，単一パス法にてクラスタリングを行う．

Step.4 年しか持たない履歴文（以下，年データ）のうち，クラスタと同年の年データを取得し，単一パス法にてクラスタリングを行う．

Step.5 クラスタに属さない年月データを取得し，各年月データを持つクラスタを作成する．

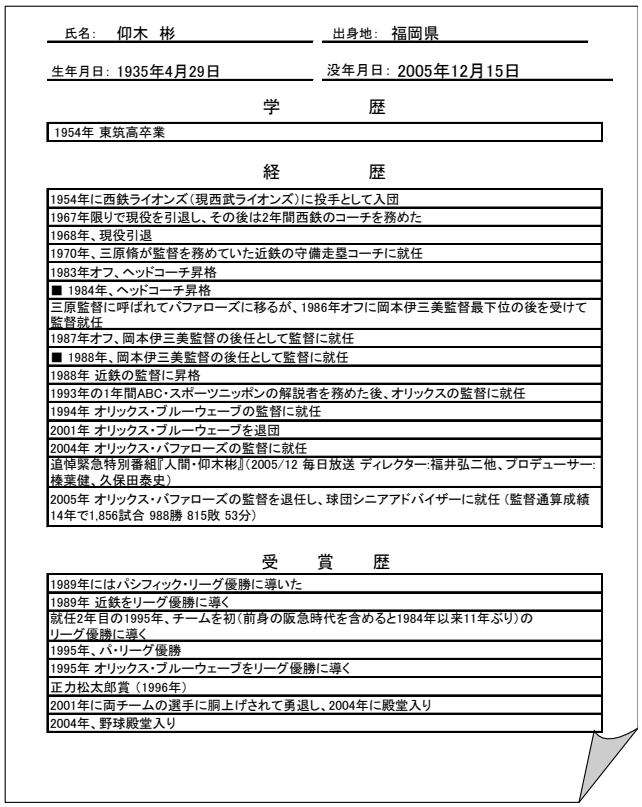
Step.6 **Step.5** で作成したクラスタのうち，同年月のクラスタを単一パス法にてクラスタリングを行う．

Step.7 **Step.5** で作成したクラスタと同年の年データを取得し，単一パス法にてクラスタリングを行う．

Step.8 クラスタに属さない年データを取得し，各年データを持つクラスタを作成する．

Step.9 **Step.8** で作成したクラスタのうち，同年のクラスタを単一パス法にてクラスタリングを行う．

図 2 時間を考慮した同義の履歴文のクラスタリング


氏名： 仰木 彬　　出身地： 福岡県

生年月日：1935年4月29日　　没年月日：2005年12月15日

学　歴

1954年 東筑高卒業

経　歴

1954年に西鉄ライオンズ（現西武ライオンズ）に投手として入団
1967年限りで現役を引退し、その後は2年間西鉄のコーチを務めた
1968年、現役引退
1970年、三原脩が監督を務めていた近鉄の守備走塁コーチに就任
1983年オフ、ヘッドコーチ昇格
■ 1984年、ヘッドコーチ昇格
三原監督に呼ばれてバファローズに移るが、1986年オフに岡本伊三美監督最下位の後を受けて監督就任
1987年オフ、岡本伊三美監督の後任として監督に就任
■ 1988年、岡本伊三美監督の後任として監督に就任
1988年 近鉄の監督に昇格
1993年の1年間ABC・スポーツニッポンの解説者を務めた後、オリックスの監督に就任
1994年 オリックス・ブルーウェーブの監督に就任
2001年 オリックス・ブルーウェーブを退団
2004年 オリックス・バファローズの監督に就任
追悼緊急特別番組『人間・仰木彬』(2005/12 毎日放送 ディレクター:福井弘二他、プロデューサー:棒葉健、久保田泰史)
2005年 オリックス・バファローズの監督を退任し、球団シニアアドバイザーに就任（監督通算成績14年で1,856試合 988勝 815敗 53分）

受　賞　歴

1989年にはパシフィック・リーグ優勝に導いた
1989年 近鉄をリーグ優勝に導く
就任2年目の1995年、チームを初（前身の阪急時代を含めると1984年以来11年ぶり）のリーグ優勝に導く
1995年、パ・リーグ優勝
1995年 オリックス・ブルーウェーブをリーグ優勝に導く
正力松太郎賞（1996年）
2001年に両チームの選手に胴上げされて勇退し、2004年に殿堂入り
2004年、野球殿堂入り

図 3 「仰木 彬」での例

日等，人生で一度しかない出来事は「戸籍」に比較的多く分類される．

3. 実　行　例

元プロ野球選手・監督の「仰木 彬」氏での，本手法を用いた例について述べる．

まず,「仰木 彬」をクエリに Google Web APIs を用いて Web ページを検索した．例では上位 50 件を取得した．得られた Web ページについて，2.2 節の§1，2 の処理を行い，211 の履歴文が得られた．211 の履歴文について，人物と履歴文の関係性判定を行い, 135 の人物と関係のある履歴文が得られた．135 の履歴文を 4 つのカテゴリに分類する．結果，戸籍に 6，学歴に 1，経歴に 21，受賞歴に 10 の履歴文が分類された．最後に，同義の履歴文のクラスタリングを行う．最終的に，戸籍に 3，学歴に 1，経歴に 16，受賞歴に 8 のクラスタとなった．

図 3 は「仰木 彬」をクエリに用いた本手法の提示例である．なお，生年月日，没年月日，出身地は「戸籍」に分類され，クラスタリングを行った履歴文に簡単なヒューリスティックを用いて，抽出したものである．抽出対象は,「1935 年 4 月 29 日，福岡県生まれ」など 3 つの履歴文を含むクラスタ,「没年月日 2005 年 12 月 15 日（満 70 歳没）」など 2 つの履歴文を含むクラスタ,「2005 年シーズン終了から僅か二ヶ月後の 12 月 15 日午後 4 時 10 分，肺癌による呼吸不全のため，福岡県内の病院で死去，70 歳没」のみ含むクラスタである．抽出された情報は，全て正しいものであった．簡単なヒューリスティックは以下の通りである．

生年月日 「生年月日」「誕生日」「生まれ」を最も多く含むクラスタの時間を抽出

没年月日 「没年月日」「死去」を最も多く含むクラスタの時間を抽出

出身地 生年月日を抽出したクラスタに最も多く含まれる都道府県を抽出

Web 検索エンジンから得られた Web ページそのままを用いて生年月日，没年月日，出身地を抽出する場合，このような簡単なヒューリスティックにより正しい情報を抽出するのは困難である．本手法では，人物に関係があるかどうかを判定し，その上でカテゴリ毎に分類，クラスタリングを行っている．そのため，簡単なヒューリスティックでも正しい情報の抽出が可能となる．本手法の処理により抽出対象が絞られているため，抽出時間も短く済む．

4. 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，人物と履歴文の関係性判定の有効性を確認するため，実験 2 では，履歴文の分類と同義の履歴文のクラスタリングについて，有効性を確認するために評価実験を行った．実験 3 では，提案手法の履歴書形式による提示方法の有用性を確認するため，被験者実験を行った．実験 4 では, Wikipedia から得る情報が提案手法に与える影響を調査した．

表 4　データセット*7

氏名	職業	履歴文
安達祐実	女優	410
宮沢りえ	女優	558
仰木彬	元プロ野球選手	193
荒川静香	プロスケーター	376
香取慎吾	タレント	514
黒住祐子	タレント	575
黒木瞳	女優	538
佐々木主浩	元プロ野球選手, 野球解説者	257
佐々木裕司	不明	509
小渕恵三	政治家	259
松坂大輔	プロ野球選手	471
青島幸男	元政治家, タレント, 作家	361
石川遼	プロゴルファー	214
川淵三郎	元プロサッカー選手	241
長嶋茂雄	元プロ野球選手	353
天海祐希	女優	421
渡辺淳一	作家	123
藤原正彦	大学教員, 作家	246
鳩山由紀夫	政治家	265
武部勤	政治家	432
北川正恭	大学教員, 元政治家	198
堀江貴文	会社役員	387
木下斉	会社役員	314
野中広務	元政治家	241
野茂英雄	元プロ野球選手	364
有森裕子	元マラソン選手	184

4·1 データセット

実験では，現代用語の基礎知識選 ユーキャン新語・流行語大賞*8 （以下，流行語大賞）にて年間大賞を受賞した 27 人のうち，提案手法にて教師データに用いた「小泉 純一郎」を除く 26 人（以下，受賞者）の氏名を用いた（表 4 参照）．

26 人の氏名を使用し，Google Web APIs を用いて Web ページを検索*9 した．その結果，得られた各上位 50 件を処理対象とした．取得した Web ページについて，2.2 節の §1，2 の処理を行い，その後，各受賞者と関係がある履歴文かどうかを人手で判定した．判定を元に各受賞者と関係のある履歴文 5935，関係のない履歴文 3069 に分類し，関係があるかどうか判定不能であった履歴文は除外した．これら 9004 の履歴文を実験に使用した (表 4 参照)．

4·2 評 価 尺 度

実験 1，2 共通の評価尺度として，適合率と再現率を用いた．適合率，再現率の定義は以下の通りである．

$$\text{適合率} = \frac{N \cap R}{N} \tag{1}$$

$$\text{再現率} = \frac{N \cap R}{R} \tag{2}$$

N，R の定義は各実験にて述べる．

4·3 実　　験　　1

人物と履歴文の関係性判定の有効性を確認するため，比較手法を用いた評価実験を行った．

§1 方　法

4.1 節で作成した 9004 の履歴文を評価用データに，9004 の履歴文のうち，受賞者と関係のある 5935 の履歴文を正解データに用いた．

比較手法として，プレーンテキストを用いた手法と，HTML の木構造を用いた手法を使用した．

プレーンテキストを用いた手法は，HTML のタグ情報を用いず，テキストのみから判定する手法である．提案手法でパターンに用いている特徴のうちタグ情報以外の特徴を用いて，非線形 SVM にて学習させた．使用する特徴は，表 1 の 1 から 4 と 20，21，23，24，26，27，29，30 である．

HTML の木構造を用いた手法は，米井らの手法 [米井 08] を応用した．米井らは構造化文書から木構造のパターンを生成する手法として k-照応組合せ部分木を提案している．米井らは，XML を対象としているが，本研究では HTML を対象とする．HTML は半構造化文書であるため，構造化を目的としないタグも存在する．そのため，木構造を解析するタグについては，構造化に使用される可能性が高いタグのみを対象とした．具体的には，2.2 節の §3 にて解析対象としたタグ（表 2 参照）と同じもの*10 を対象とした．対象のタグについて，k-照応組合せ部分木を用いてパターンを生成した．k の値については，k=∞とした．得られたパターンを非線形 SVM にて学習させた．

プレーンテキストを用いた手法，HTML の木構造を用いた手法，どちらも教師データには提案手法と同じものを用いた．

各手法の適合率，再現率，F 値を比較した．適合率，再現率における N は各手法にて受賞者と関係があると判定

*7　「履歴文」の数は, 人物と関係があるか判定不能であった履歴文を除去した．
*8　http://singo.jiyu.co.jp/
*9　検索の結果，同姓同名人物が含まれる可能性がある場合，人物に関係の深いキーワード（例えば所属など）を用い AND 検索を行った．
*10　br と hr について空要素のため対象から外している．

された履歴文の数，R は人手により受賞者と関係があると判定した履歴文の数である．F 値は，

$$F\text{ 値} = \frac{2 \cdot \text{適合率} \cdot \text{再現率}}{\text{適合率} + \text{再現率}} \quad (3)$$

である．なお，後述の集計結果は適合率，再現率，F 値いずれも，各受賞者の値を平均した値である．

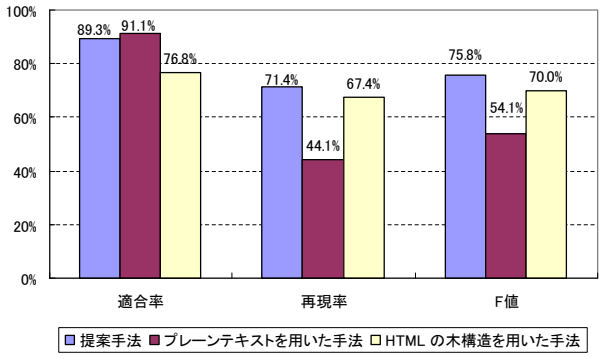


図 4 人物と履歴文の関係性判定性能

§2 結果と考察

適合率は，プレーンテキストを用いた手法が最も良かった（91.1%）．再現率，F 値については，提案手法が最も良かった（再現率：71.4%，F 値：75.8%）．提案手法は，適合率でプレーンテキストを用いた手法には若干劣るが 9 割近い評価（89.3%）であった．また，再現率は 3 手法のうち最も高い評価で，7 割を超える正解の履歴文を判定できた．適合率と再現率の総合的な評価尺度である F 値も 3 手法では最高であり，提案手法の有効性を示す結果であると考える（図 4 参照）．

実験に用いた教師データを評価データとし，正解データを教師データとして交差検定を行った．その結果，提案手法は適合率 80.7%，再現率 92.6%，F 値 86.3%，プレーンテキストを用いた手法は適合率 87.0%，再現率 61.9%，F 値 72.4%，HTML の木構造を用いた手法は適合率 71.0%，再現率 75.2%，F 値 73.1%であった．適合率は，プレーンテキストを用いた手法が，再現率，F 値は，提案手法が最も良かった．この結果は，実験結果と同じ傾向であり，データ集合の偏りが実験結果に影響していないことを示唆するものであると考える．

プレーンテキストを用いた手法にて受賞者と関係があると判定された履歴文の多くは，受賞者の氏名または姓を含むものであった．提案手法では，HTML のタグ情報もパターンとして用いているため，受賞者の氏名または姓を含まないが受賞者と関係のある履歴文を数多く判定できた．

HTML の木構造を用いた手法は，HTML 全体の論理構造を解析する必要があり HTML にタグの記述ミスが含まれる場合，論理構造を解析に失敗する可能性がある．実験でも，パターンの生成失敗例がいくつか見られた．

提案手法では，主に HTML のタグの出現パターンを用いているため，HTML 全体の論理構造の解析は必要ない．実験では，全ての履歴文でパターンの生成に成功した．また，HTML の木構造を用いた手法では，抽出した履歴文の周辺の文字列の情報が利用されていない．提案手法では，履歴文の周辺の文字列も利用するため，適合率が高くなったと考える．

4·4 実　　験　　2

履歴文の分類と同義の履歴文のクラスタリングについての有効性を確認するため，分類性能とクラスタリング性能についての評価実験を行った．

§1 方　法

2 つの評価用データを作成した．1 つは，4.1 節で人手により判定した受賞者と関係のある履歴文（以下，人手判定データ）5935 件である．もう 1 つは，データセットを用いて，2.2 節の §3 の処理により受賞者と関係があると判定された履歴文（以下，提案手法データ）4516 件である．この 2 つの評価用データを履歴文の分類の評価に用いた．

人手判定データについて，履歴文を 2.3 節にて定義した 4 つのカテゴリ（戸籍，学歴，経歴，受賞歴）のどれか 1 つに人手で分類し，履歴文の分類の正解データとした．どのカテゴリにも合致しないものは省いた．

履歴文の分類の評価には，適合率と再現率を用いた．N は分類処理により分類された各カテゴリ内の履歴文の数，R は N に対応する正解データ内の履歴文の数である．なお，集計結果は 4 つのカテゴリの結果を平均した値である．

同義の履歴文のクラスタリングの評価用データは，上述した履歴文の分類の正解データである．履歴文の分類の正解データを人手で同義の履歴文毎に分類し，同義の履歴文のクラスタリングの正解データとした．クラスタリング手法内で用いる単一パス法の閾値は，予備実験にて相対的に評価が良かった 0.5 とした．

同義の履歴文のクラスタリングの評価では，適合率と再現率に加え，どれだけ履歴文を削減できたかの指標として圧縮率も併せて計算した．適合率と再現率における N はクラスタリング処理により得られた各クラスタ内の履歴文の数，R は N に対応する正解データ内の履歴文の数である．圧縮率は，

$$\text{圧縮率} = 1 - (C/I) \quad (4)$$

と定義した．C は作成されたクラスタの数，I は入力した履歴文の数である．なお，集計結果はカテゴリ毎に平均した値である．

§2 結果と考察

人手判定データに対する分類性能は，戸籍，学歴，受賞歴について，いずれも適合率が 9 割以上であった．経歴については，適合率 61.9%，再現率 55.7%と他の 3 つと比べ，評価が低かった（表 5 参照）．

戸籍，学歴，経歴，受賞歴の教師データと人手判定データを用いて 5 分割交差検定を行った．その結果，戸籍で適合率 86.0%，再現率 90.1%，学歴で適合率 92.1%，再現率 94.3%，経歴で適合率 82.8%，再現率 82.7%，受賞歴で適合率 80.0%，再現率 81.5%（全て平均）と，いずれも 8 割以上であった．

提案手法データに対する分類性能は，適合率では人手判定データと同等であった．再現率は，1〜2 割程度，評価が低下したが経歴以外は 5 割以上であった．

流行語大賞の受賞者の職業は様々であり，職種の違いにより評価が異なることが考えられる．そのため，職種別に受賞者を分類し再集計を行った．分類した職種は，芸能人（俳優，女優，タレントなど），スポーツ選手（過去にスポーツ選手だった人物も含む），政治家（過去に政治家だった人物も含む），その他（芸能人，スポーツ選手，政治家のどれにも当てはまらない人物）である．その他には，作家，大学教授，会社役員などが分類されている．なお，複数に当てはまる人物は最も有名である職種に分類した．

集計の結果，戸籍，学歴，受賞歴の適合率については，職種に関係なく高い結果が得られた．再現率については，人手判定データの芸能人，政治家，その他で 7 割を超えたがスポーツ選手については若干評価が低下した．経歴については，政治家，その他で適合率 76%，スポーツ選手で 65%であったのに対し，芸能人では，35%と大幅に低下した（表 6 参照）．

提案手法データは，自動で人物と関係があるかどうかを判定したデータである．そのため，必ずしも分類の正解データが提案手法データの中にあるとは限らない．同時に，人物と関係ないデータも含まれている．このような不完全なデータであるにも関わらず，適合率の評価は人手判定データと同等であり，再現率も平均 5 割を超えた．この結果は，ある程度間違ったデータを与えても頑健に動作することを示唆していると考える．同時に，実験 1 にて評価を行った提案手法の人物と履歴文の関係性判定の有効性を裏付ける結果でもあると考える．

職種別の集計では経歴で違いが見られた（表 6）．政治家やその他（特に大学教授，会社役員など）において，企業や団体に所属し，役職がある職業の人物で良い結果であった．一方，芸能人は，適合率，再現率ともに他の職種より低かった．理由として，芸能人は，企業や団体に所属しない場合があることや，所属していてもあまり Web 上で触れられないことから，本研究での経歴の定義に当てはまる情報が得られないことがあげられる．また，役職や企業名が配役の記述として含まれる履歴文（例えば，「1999.09.18 金融腐食列島呪縛」 東映＝角川書店＝産經新聞社... 佐藤弘子秘書」）が誤って経歴に分類されることが別の理由である．提案手法の経歴の分類は，所属企業，団体や役職がある人物について特に有効であると言える．

クラスタリング性能の評価については，戸籍，学歴，経歴で適合率 100.0%であった．平均でも 99.8%であり，非常に良い結果であった．再現率は，平均 84.0%で，4 つのカテゴリ全てで 8 割を超える評価であった．圧縮率は平均 38.3%であった（表 7 参照）．適合率，再現率ともに非常に高い評価であるため，クラスタ内にまとめられた履歴文の大部分は重複した内容であることがわかる．提案手法のクラスタリングにより，重複した履歴文を平均約 4 割削減できた．同じ意味の文を複数閲覧させることは，利用者にとって負担となる．提案手法のクラスタリングにより，利用者の履歴書閲覧の負担を軽減できると考える．

表 5 分類性能

	人手判定データ		提案手法データ	
	適合率	再現率	適合率	再現率
戸籍	92.4%	80.0%	89.2%	58.0%
学歴	94.7%	77.7%	90.0%	65.7%
経歴	61.9%	55.7%	62.5%	43.4%
受賞歴	93.3%	69.8%	91.1%	53.9%

表 6 職種別分類性能

		人手判定データ		提案手法データ	
		適合率	再現率	適合率	再現率
芸能人 (6 人)	戸籍	81.3%	77.3%	81.0%	54.0%
	学歴	100.0%	91.7%	100.0%	91.7%
	経歴	35.1%	16.0%	40.0%	11.8%
	受賞歴	96.3%	78.2%	96.7%	50.9%
スポーツ選手 (9 人)	戸籍	98.2%	84.1%	94.1%	60.4%
	学歴	90.0%	66.3%	75.0%	45.6%
	経歴	65.3%	56.0%	67.2%	47.5%
	受賞歴	88.9%	54.4%	87.1%	42.4%
政治家 (6 人)	戸籍	93.6%	83.6%	94.1%	64.7%
	学歴	100.0%	87.5%	100.0%	70.8%
	経歴	76.6%	89.3%	66.7%	63.8%
	受賞歴	95.8%	84.2%	100.0%	81.2%
その他 (5 人)	戸籍	84.4%	74.1%	83.3%	50.2%
	学歴	92.5%	79.6%	90.0%	64.6%
	経歴	76.2%	69.6%	69.5%	49.6%
	受賞歴	96.0%	70.1%	86.0%	45.3%

表 7 クラスタリング性能

	平均	戸籍	学歴	経歴	受賞歴
適合率	99.8%	100.0%	100.0%	100.0%	99.2%
再現率	84.0%	80.8%	91.3%	83.1%	85.9%
圧縮率	38.3%	62.1%	10.5%	31.4%	28.7%

4·5 実　　験　　3

提案手法の履歴書形式による提示方法の有用性の確認のため，被験者を用いた実験を行った．

§1 方 法

実験 2 で職種別に分類した受賞者の中から 3 人ずつ合計 12 人の受賞者を抽出した.

比較する提示手法として 2 種類用意した．1 つは，2.2 節の処理で得た履歴文を時系列に並べるもの（以下，提示手法 1），もう 1 つは，2.2 節で得た履歴文を，クラスタリングせずに 2.3 節の処理で戸籍，学歴，経歴，受賞歴のカテゴリ毎に提示するもの（以下，提示手法 2）である．提案手法では，3 章で説明したとおり, 生年月日, 没年月日，出身地を自動で抽出[*11] して提示した.

問題は 2 種類用意した．1 つは，ある時間に起った出来事について解答させる問題である（以下，出来事への解答）．例えば,「2006 年に入学した学校は何ですか？」である．もう 1 つは，ある出来事について起った時間を解答させる問題である（以下，時間への解答）．時間への解答では，ある出来事を没年月日とした.

被験者は情報系大学院生 9 人である．提示手法 1，2, 提案手法の解答が 4 つずつ得られるように被験者に問題を割り当てた．2 種類の問題を各受賞者につき 1 つずつ割り当てた.

実験用評価システムを作成した. 被験者に提示手法 1, 2，提案手法のいずれかで作成した画面を提示し, 問題に答えさせ, 解答時間を計測した.

実験後, 各提示手法で作成された画面が履歴書に見えるかどうかと，最もわかりやすかった提示方法を質問紙にて答えさせた.

§2 結果と考察

解答時間について集計したところ，出来事への解答・時間への解答，いずれも提案手法が最も短かった．正答率ついては，出来事への解答・時間への解答の違いに関わらず，いずれの手法も結果が良かった（表 8 参照）.

解答時間について一要因の分散分析による検定を行った．分析では、出来事への解答・時間への解答を個別データとして扱った．結果，各手法間に有意差が見られた（$F = 4.14 > F_{\alpha=0.025}(2, 213) = 3.80, p < .05$）．Fisher の LSD 法を用いて下位検定を行ったところ，提案手法と提示手法 1 の間に有意差が見られた（$p < .01$）．正答率について一要因の分散分析を行ったところ有意差は見られなかった（$F(2, 35) = 2.01$）．上記の結果より, 解答時間については，履歴文を単純に時系列で並べるよりも，履歴書形式で提示することの有効性を確認できたと考える. 正答率については, 提案手法では, 履歴文を最も加工, 圧縮しているにも関わらず，他手法と同等の結果が得られることを示していると考える.

質問紙への回答についても集計した．履歴書に見えるかについては，提示手法 1 で 0.0%(0/9)，提示手法 2 で 88.9%(8/9)，提案手法で 100.0%(9/9) であった．最もわかりやすかった提示方法については，提示手法 1 で 0.0%(0/9), 提示手法 2 で 22.2%(2/9) , 提案手法で 77.8%(7/9) であった（表 9 参照）.

解答時間は提案手法が最も短く，9 割以上の正答率が得られた．また，被験者全てが提案手法が履歴書に見えると回答し，最もわかりやすい提示方法として提案手法を挙げた被験者が最も多かった．これらの結果より，提案手法の履歴書形式による提示方法の一定の有用性が確認できたと考える.

表 8 解答時間と正答率

		解答時間	正答率
全体 [*12]	提示手法 1	55.33 秒	91.7%
	提示手法 2	46.18 秒	98.6%
	提案手法	34.53 秒	97.2%
出来事への解答	提示手法 1	56.97 秒	86.1%
	提示手法 2	52.14 秒	100.0%
	提案手法	38.94 秒	100.0%
時間への解答	提示手法 1	53.69 秒	97.2%
	提示手法 2	40.22 秒	97.2%
	提案手法	30.11 秒	94.4%

表 9 質問紙による回答結果

	履歴書に見えるか？	最もわかりやすかった提示方法
提示手法 1	0.0%	0.0%
提示手法 2	88.9%	22.2%
提案手法	100.0%	77.8%

4·6 実 験 4

提案手法は，特定のサイトから得られる Web ページに限定した手法ではないが，人物について詳細な記述がされる百科事典サイトやデータベースサイトから得られる情報に影響を受ける可能性がある．特に，Wikipedia の影響が大きいことが考えられる. 本節では，Wikipedia から得る情報が提案手法に与える影響を調査した.

§1 データセット

データセット中の履歴文について，Wikipedia から得られた割合を調査したところ，平均 23.4%であった．人物により 0.0%から 50.7%と割合に幅があった．集計結果を表 10 に示す.

§2 人物と履歴文の関係性判定

人物と履歴文の関係性判定の教師データについて, Wikipedia から得た履歴文は 9.3%(627/6740) であった．627 の履歴文の 88.0%(552/627) は人物と関係があると判定された履歴文（正解データ）であり，教師データの正解データに占める割合は 24.3%である．人物と履歴文の関係性判定について，Wikipedia から得る情報がどの程度評価に影響を与えているか調査を行った．4.3 節における提案手法の結果について，Wikipedia から得た履歴文を除外して再集計を行った．その結果，適合率 86.5%，再現率 68.0%，F 値 72.3%と, Wikipedia から得た履歴文を含めた評価（適

[*11] 抽出された生年月日，没年月日，出身地は，全て正しいものであった.

[*12] 出来事への解答，時間への解答を合わせて集計した結果.

表 10 データセットにおける Wikipedia の割合

	Wikipedia	Wikipedia の割合
安達祐実	168	41.0%
宮沢りえ	149	26.7%
仰木彬	47	24.4%
荒川静香	185	49.2%
香取慎吾	118	23.0%
黒住祐子	129	22.4%
黒木瞳	147	27.3%
佐々木主浩	57	22.2%
佐々木裕司	0	0.0%
小渕恵三	89	34.4%
松坂大輔	61	13.0%
青島幸男	57	15.8%
石川遼	25	11.7%
川淵三郎	68	28.2%
長嶋茂雄	179	50.7%
天海祐希	110	26.1%
渡辺淳一	38	30.9%
藤原正彦	44	17.9%
鳩山由紀夫	44	16.6%
武部勤	56	13.0%
北川正恭	42	21.2%
堀江貴文	37	9.6%
木下斉	40	12.7%
野中広務	53	22.0%
野茂英雄	87	23.9%
有森裕子	46	25.0%
平均	79.8	23.4%

「Wikipedia」は Wikipedia サイト内から得た履歴文の数.

合率 89.3%，再現率 71.4%，F 値 75.8%）より，若干低下した. 各尺度とも数%の低下にとどまり，Wikipedia から得た履歴文が与える提案手法への影響は小さいと考える.

§3 履歴書形式で提示される履歴文

履歴書形式で提示される履歴文について, Wikipedia から得た情報の影響を調査した．4.4 節で用いた提案手法データと，Wikipedia から得た履歴文を除外した提案手法データ（以下，Wikipedia 除外データ）について 2.4 節の処理を行ったものを使用した．2 つのデータを用いて，Wikipedia を用いない場合の提示割合を集計した．集計は，全体と，4.4 節§2 で定義した職種（芸能人，スポーツ選手，政治家，その他）について行った．以下の尺度を用いた.

$$\text{提示割合} = \frac{Wikipedia \text{ 除外データから 2.4 節の処理にて得られたクラスタ数}}{\text{提案手法データから 2.4 節の処理にて得られたクラスタ数}} \quad (5)$$

Wikipedia 除外データを使用しても，平均 61.3%（表 11 参照）提示できた．この結果は，Wikipedia から得た情報を用いなくても，ある程度の情報を提示できることを示している．特に戸籍，経歴は 7 割以上提示できた．一方で，学歴，受賞歴はどちらも約 5 割であり，戸籍，経歴よりも Wikipedia の情報への依存度が大きいことがわかった．職種別の集計の結果，芸能人が最も Wikipedia から影響を受け，政治家があまり影響を受けないことがわかった．芸能人は，他の職種の人物よりも Wikipedia で詳細に記述されるからであろう．政治家は，自身のサイトなどで詳細なプロフィール情報を公開していることが多く，Wikipedia の情報に頼らなくても戸籍，学歴，経歴，受賞歴の各情報を得られやすいためであると考える.

表 11 履歴書形式で提示される履歴文における Wikipedia の影響

	平均	戸籍	学歴	経歴	受賞歴
全体	61.3%	78.6%	50.9%	73.2%	42.6%
芸能人	48.9%	74.7%	50.0%	41.7%	29.2%
スポーツ選手	55.0%	79.2%	25.0%	76.4%	39.3%
政治家	73.5%	84.2%	79.2%	79.7%	51.1%
その他	61.5%	75.0%	48.1%	75.3%	47.5%

5. 関連研究と議論

5·1 関 連 研 究

本研究に類似する研究として，経歴を抽出する研究 [Kimura 07] や伝記を作成する研究 [Kim 02, Schiffman 01] がある．[Kimura 07] では，得られた Web ページを同姓同名人物毎にクラスタリングを行い，人物に関する経歴を抽出し，時系列に並べて提示している．本研究とは，以下の点で異なる．まず，提示方法に大きな違いがある．[Kimura 07] は，得られた経歴のカテゴリ毎の分類は行わず，単純に時系列に並べて提示している．本研究は，戸籍，学歴，経歴，受賞歴の 4 つのカテゴリ毎に分類し，履歴書の形式にあわせて提示する．履歴書という多くの人になじみのある形式で提示したほうが，利用者にとって理解がしやすくなると考える．次に，人物との関係性判定方法が異なる．[Kimura 07] では，[米井 08] と同じく人物と Web 上から得られた経歴の判定に木構造をパターンに用いて判定を行っている．HTML で記述された Web ページは雑多であり，記述ミスがある可能性もあるため，木構造を用いた手法では有効に機能しない場合がある．本研究では，木構造ではなく，主に HTML のタグの出現パターンを用いて人物との判定を行っている．[Kim 02] ではニュース記事を，[Schiffman 01] では Web 上の情報を用いて，伝記を自動的に生成している．本研究は，Web 上の情報を用いるという点で [Schiffman 01] と類似する．提案手法は，伝記の自動作成にも応用可能であると考える.

本研究の人物と履歴文の関係性判定は，オブジェクト同士の関連性を求める研究と類似する．オブジェクト同士の関連性を求める研究には，米井ら [米井 08]，大前ら [大前 06]，Yoshida ら [Yoshida 04] の研究がある．米井らの研究は，構造化文書から木構造のパターンを生成している．本研究では，半構造化文書である HTML を対象とし，全体の論理構造の解析を必要としない手法を提案した．大前ら，Yoshida らの研究は，HTML の表 (table) から属性と属性値の関連性を判定している．本研究の人物と履歴文の関係性判定は，表 (table) 以外のタグも考慮している.

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 [森 05]，呼称や別名を抽出す

る研究 [外間 06, 若木 08, 本間 07]，肩書や職業に関連する情報を抽出する研究 [Wan 05, 上田 09] などがある．[森 05] では，語の共起を利用してキーワードを抽出している．[外間 06, 若木 08, 本間 07] は，人物に対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．[本間 07] は，パターンの自動生成を同時に行っている．[Wan 05, 上田 09] は，同姓同名人物毎に分離された Web ページクラスタにラベル付けすることを目的に，肩書や職業に関連する情報を抽出している．本研究では，時間と人物に関する出来事の両方を含む文（履歴文）を抽出した上で，戸籍，学歴，経歴，受賞歴のカテゴリ毎に分類している．

Web 上の同姓同名人物の分離についての研究には，佐藤ら [佐藤 05]，Bekkerman ら [Bekkerman 05] ，白砂ら [白砂 06]，木村ら [木村 06] の研究などがある．佐藤ら，Bekkerman ら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，プロフィール情報に，木村らは Web 検索エンジンから得られるスニペットに着目した手法を提案している．本研究では，同姓同名人物毎に分離された Web ページの使用を対象としている．

5·2 議　　論

本研究では，Web 人名検索で検索される人物の理解を支援するために，社会で広く人物の理解のために用いられる履歴書に着目し, Web ページから時間と人物に関する出来事の両方を含む文（履歴文）を抽出して履歴書の形式で提示する手法を提案した．本研究の意義は，Web ページからの人物に関連する情報抽出だけではなく，抽出した情報の統合・提示方法についても焦点をあてた点にある．

評価実験の結果，以下の知見を得た．

(1) 提案手法の人物と履歴文の関係性判定の判定性能は，適合率 89.3%，再現率 71.4%，F 値 75.8%であった．再現率，F 値については，比較手法の中では最も良い結果であった．適合率も 9 割近い評価であり，これらは提案手法の人物と履歴文の関係性判定の有効性を示す結果であると考える．

(2) 分類性能は，人手判定データの評価について 3 つのカテゴリで適合率が 9 割以上であった（戸籍：92.4%，学歴:94.7%，受賞歴:93.3%）．再現率は，平均 70.8%であった．また，多少の間違いを含む提案手法の人物と履歴文の関係性判定により得られた履歴文であっても，人手で人物と関係があると判定した履歴文を分類した結果と同等の適合率が得られた．この結果は，ある程度間違ったデータを与えても頑健に動作することを示唆していると考える．同時に，提案手法の人物と履歴文の関係性判定の有効性を裏付ける結果でもあると考える．クラスタリング性能は，適合率 99.8%，再現率 84.0%であり，非常に良い結果であった．この結果は，提案手法の時間を考慮したクラスタリングの有効性を示す結果であると考える．

(3) 被験者を用いた提案手法と 2 つの比較手法の提示方法についての比較実験では，提案手法の解答時間が最も短く，正答率も良かった（出来事への解答：100.0%，時間への解答：94.4%）．また，最もわかりやすい提示方法として提案手法を選ぶ被験者が最も多かった．これらの結果より，提案手法の履歴書形式による提示方法の一定の有用性を確認できたと考える．

(4) Wikipedia から得る情報の影響について調査した．人物と履歴文の関係性判定への影響は小さいことがわかった．また，履歴書形式で提示される履歴文（クラスタ）について, Wikipedia から得る情報がなくても, 平均 61.3%の履歴文を提示できることがわかった．この結果は，Wikipedia から得る情報がなくても，ある程度の情報を提示できることが示唆されたものと考える．

以下では，提案手法における個別手法の一般性と，提案手法と個別手法に共通する限界について述べる．

提案手法における履歴文の抽出は，人名以外のクエリでも利用可能である．例えば，クエリに団体名や企業名を用いた場合，団体や企業の歴史に関する文の抽出が期待でき, クエリについての年表作成に応用が可能である．また，人物と履歴文の関係性判定は，クエリと Web ページ内の単文以外のオブジェクトとの関係性も判定可能である．Web 上でのクエリとオブジェクトの関係性の判定は重要な課題であり，研究 [米井 08, 大前 06, Yoshida 04] も多数行われている．人物と履歴文の関係性判定の応用範囲は非常に広いと考える．

同義の履歴文のクラスタリングは時間を考慮しているが，時間以外にも適応可能である．例えば，時間を金額に置き換えて，製品とその値段を含む文のクラスタリングに応用できる．具体的には，「ソニー・コンピュータエンタテインメント PLAYSTATION 2 ￥14,800(税込)」,「ソニー・コンピュータエンタテインメント PLAYSTATION3 ￥39,980(税込)」の 2 つの文が得られたとする．これら 2 つの文は，別の製品について言及されているが文字列の構成がほとんど同じであるため，一般的なクラスタリングでは，同じクラスタとしてまとめられる可能性が高い．金額を考慮したクラスタリングを行えば別のクラスタに分類できる．

提案手法（履歴文の抽出，分類，クラスタリング）では，得られた情報の取捨選択や統合処理を行っているため，ある程度の数の Web ページを必要とする．Web 上での言及が少ない人物（人名）の場合，得られる Web ペー

ジの数が少なくなり，各手法にて取捨選択や統合処理を行うと，提示される情報がほとんどないことがありうる．提案手法は，少量の Web ページしか得られない人名については改良が必要である．

以下に今後の課題をまとめる．

人物と履歴文の関係性判定の再現率は 71.4%と良い結果が得られたが，3 割弱については判定に失敗した．判定に失敗したものは，パターン作成に用いた Web ページのパターンに当てはまらないものが多かった．これは，教師データの正解データの約 1/4 が Wikipedia から得た情報と，得られたサイトに若干の偏りがあったことが影響を与えたと考える．改善策として，Wikipedia 以外のサイトのパターンを幅広く収集することが挙げられる．

履歴文の分類については，経歴の分類性能が他のカテゴリと比べ評価が低下した．特に，芸能人については顕著であった（適合率：35.1%，再現率：16.0%）．本研究では，人物の企業や団体への就職，加入，参加の記述を経歴として分類している．芸能人の場合，所属事務所等の加入に関する記述が本研究での経歴となる．しかし，芸能人についての所属事務所等の加入と加入時期を合わせて記述されることはほとんどない．実験中に用いた芸能人のうち,「安達祐実」,「宮沢りえ」の 2 人については，分類の正解データ作成の際，本研究で定義する経歴に合致するものは得られなかった．一方で，芸能人については本研究における経歴に合致しないものが分類されやすい傾向にあった．ここで，芸能人 6 人の経歴に分類された履歴文を調査した．その結果，ほとんどが映画，テレビ番組，舞台などの出演歴であった．特に，企業や団体の名前が入った出演歴が多かった．例えば,「安達祐実」の「ハウス食品「ふっくらケーキ屋さん」（1995 年 - 1996 年)」などが該当する.「ハウス食品」は企業名である．これら出演歴については，芸能人の経歴とされることがある．プロフィールにも，出演歴が記載される場合が多い．ここで，芸能人のみ出演歴も経歴とした場合の適合率を集計した．その結果，人手判定データで 90.0%，提案手法データでは 100.0%であった．このように，職種によって何を経歴とするかには揺れがあり，経歴の定義によっては，評価に大きな差が生まれる．今後は，職種別に何を経歴とするかについて検討を行った上で，職種別に経歴の分類パターンを変更する改善を行いたい．

同義の履歴文のクラスタリングに対する評価は，適合率，再現率ともに高く，同義の履歴文の多くを 1 つのクラスタにまとめられることが確認できた．しかし，図 3 の経歴に分類されている「1983 年オフ，ヘッドコーチ昇格」「■ 1984 年，ヘッドコーチ昇格」のように，同義であっても記述されている時間が異なっている場合は提案手法のクラスタリングでは対応できない．前者は，契約を行った年，後者は，就任した年であると考える．このように，書き手の解釈により，同義であっても記述される時間が異なる場合がある．同義で異なる時間が記述された履歴文をどのように扱うかについては今後の課題とする．

本研究では，戸籍，学歴，経歴，受賞歴のいずれにも分類されなかった履歴文を提示の対象としていない．しかし，分類されなかったものの中にも，人物の理解に有益なものが存在すると考える．ただし，分類．されなかったもの全てを提示すると，閲覧させる数が増え利用者の負担が増える可能性がある．利用者への負担を軽減させるために，人物の理解に有益である履歴文のランキングを行い上位数件のみ提示するなど，提示方法に配慮した手法を検討したい．

6. お わ り に

Web 人名検索で検索される人物の理解を支援するために，Web ページから時間と人物に関する出来事の両方を含む文（履歴文）を履歴書の形式で提示する手法を提案した．提案手法の特徴は, ヒューリスティックを用いた履歴文の抽出とそのフィルタリング，主に HTML のタグの出現パターンを用いた SVM による人物と履歴文の関係性判定，SVM を用いた履歴文のカテゴリへの分類，記述された時間を考慮した同義の履歴文のクラスタリングである．

本研究の意義は，Web ページからの人物に関連する情報抽出だけではなく，抽出した情報の統合・提示方法についても焦点をあてた点にある．評価実験の結果, 提案手法の有効性を確認した．履歴文の抽出は人名以外のクエリに, 同義の履歴文のクラスタリングは時間以外にも応用可能である．

今後の課題として, 人物と履歴文の関係性判定の性能向上, 芸能人における経歴カテゴリの分類性能向上, 時間が異なる同義の履歴文のクラスタリング性能向上などがあげられる．

◇ 参 考 文 献 ◇

[Bekkerman 05] Bekkerman, R. and McCallum, A.: Disambiguating Web Appearances of People in a Social Network, in *Proceedings of the 14th World Wide Web Conference(WWW2005)* (2005)

[Guha 04] Guha, R. V. and Garg, A.: Disambiguating People in Search, in *Standord University* (2004)

[外間 06] 外間 智子, 北川 博：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)

[本間 07] 本間 大輝, Bollegala, D., 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)

[Kim 02] Kim, S., Alani, H., Hall, W., Lewis, P., Millard, D., Shadbolt, N., and Weal, M.: Artequakt: Generating Tailored Biographies from Automatically Annotated Fragments from the Web, in *Proceedings of Workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM'02)*, pp. 1–6 (2002)

[木村 06] 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)

[Kimura 07] Kimura, R., Oyama, S., Toda, H., and Tanaka, K.: Creating Personal Histories from the Web using Namesake Disambigua-

tion and Event Extraction, in *Proceedings of the 7th International Conference on Web Engineering(ICWE2007), LNCS4607*, pp. 400–414 (2007)
[森 05]　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)
[大前 06]　大前 信弘, 黄瀬 浩一：Web の表を対象とした属性の自動識別, 情報処理学会研究報告, NL-171, pp. 43–48 (2006)
[佐藤 05]　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, No. SIG8, pp. 26–36 (2005)
[Schiffman 01]　Schiffman, B., Mani, I., and Concepcion, K. J.: Producing Biographical Summaries: Combining Linguistic Knowledge with Corpus Statistics, in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-EACL 2001)* (2001)
[白砂 06]　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)
[上田 09]　上田 洋, 村上 晴美, 辰巳 昭治：Web 上の同姓同名人物識別のための職業関連情報の抽出, システム制御情報学会論文誌, Vol. 22, No. 6, pp. 229–240 (2009)
[若木 08]　若木 裕美, 藤井 寛子, 福井 美佳, 住田 一男：Web 情報を用いた人物の愛称抽出, 日本データベース学会論文誌, Vol. 7, No. 1, pp. 169–174 (2008)
[Wan 05]　Wan, X., Gao, J., Li, M., and Ding, B.: Person Resolution in Person Search Results: WebHawk, in *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 163–170 (2005)
[米井 08]　米井 由美, 岩井原 瑞穂, 吉川正俊：XML 文書における構造の素性を用いた照応による人物検索, 日本データベース学会論文誌, Vol. 7, No. 1, pp. 151–156 (2008)
[Yoshida 04]　Yoshida, M., Torisawa, K., and Tsujii, J.: Integrating Tables on the World Wide Web, 人工知能学会論文誌, Vol. 19, No. 6, pp. 548–560 (2004)

〔担当委員：山田 和明〕

2009 年 4 月 30 日　受理

著　者　紹　介

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, システム制御情報学会, 日本図書館情報学会の会員.

村上　晴美（正会員）

大阪市立大学大学院創造都市研究科教授. 京都大学文学部哲学科心理学専攻, 富士通株式会社, 英国 UMIST 計算機学科修士課程, 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, 大阪市立大学学術情報総合センター講師, 助教授などを経て現職. 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. 日本認知科学会, 情報処理学会, 日本図書館情報学会, ACM などの会員.

辰巳　昭治（正会員）

1970 年大阪大学工学部通信工学科卒業. 1972 年同大学大学院工学研究科通信工学専攻修士課程修了. 同年川崎重工業（株）入社. 1978 年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, IEEE, ACM などの会員.