知能と複雑系 119−24
(2000. 1. 13)

ワールドワイドウェブからの人物情報の自動収集

山本 あゆみ　佐藤 理史

北陸先端科学技術大学院大学情報科学研究科

〒923-1292 石川県能美郡辰口町旭台 1-1
ayamamot@jaist.ac.jp　sato@jaist.ac.jp

あらまし　本稿では，ワールドワイドウェブから人物に関する情報を収集する２つの方法を提案する．第１の方法は，表形式の職業別人名リストを情報源として利用する方法である．この方法では，まず，与えられた職名（例えば「政治家」）から，検索エンジンとハイパーリンクを用いて，その職業の人名リストを収集する．次に，収集されたリストに対して表解析を適用し，それぞれの人物に対して主要情報を抽出する．第２の方法は，人物を紹介した短いテキスト（プロフィール）を抽出する方法である．この方法は，職名と人名を入力とし，それらを用いて収集したウェブページに対してレイアウト解析を適用し，求める人物のプロフィールを抽出する．

キーワード　ワールドワイドウェブ，人物情報の自動抽出，表解析，情報抽出，検索エンジン

Automatic Collection of People's Information from the World Wide Web

Ayumi YAMAMOTO　Satoshi SATO

School of Information Science,
Japan Advanced Institute of Science and Technology

Asahidai 1-1, Tatsunokuchi, Nomi, Ishikawa, 923-1292, Japan
ayamamot@jaist.ac.jp　sato@jaist.ac.jp

Abstract　This paper proposes two methods for collecting people's information from the World Wide Web. From the given occupation category such as Seijika (politicians), the first method collects web pages that include tables whose content is people lists of the given occupation, and extract personal properties such as name and birthday for each person by using table analysis. The second method accepts a person name and her occupation as an input, and collects her profile in text form by using layout analysis of HTML texts.

key words　World Wide Web, automatic extraction of people's information, table analysis, information extraction, search engine

1. はじめに

World Wide Web(WWW)には，様々な情報が膨大に存在し，これらの情報を利用するために，WWWの情報検索サービスの開発が行われてきた．その代表的なものの１つに，ロボット型検索エンジンと呼ばれる汎用の検索サービスがある．この検索サービスは，WWWから自動収集した全てのページを検索対象とする．このため，多くの場合，ある検索質問（クエリ）に対して，大量の検索結果（URL）が得られ，それらの大部分が，求める情報とはあまり関係がないものであることが多い．これでは，利用者が効率良く情報収集するのは，困難である．

一方，Yahoo!* のように人手で整理された情報を扱う検索サービスは，絞り込まれた結果を得ることができる．そこで，WWWの情報検索を効率良くする方法として，検索対象のカテゴリを限定した検索サービスが考えられる．検索対象カテゴリを限定することにより，異なるカテゴリを排除することが可能となるとともに，主要な情報がそのカテゴリによって定まるため，整理された情報を利用者に提供することが可能となる．

本研究では，検索対象を人物とし，人名からその人物の主要な情報を提供する人物情報検索サービスの実現を目的とする．我々は，これを実現するために，検索対象となる人物情報を格納したデータベースを自動生成し，これを検索時に利用するアプローチをとる．既に，人物情報データベースを自動生成する１つの方法として，WWW上に存在する職業別人名リストを利用して人名を収集する方法を提案し，「政治家」を対象とした実験を行った[1].

本稿では，職業別人名リストを利用して，人名だけでなく，他の情報（生年月日など）も自動収集する方法を提案する．さらに，人物情報のもう１つの自動収集の方法として，人名を入力としてその人物のプロフィールを自動収集する方法について述べる．

2. 人物情報自動収集システムの概要

人物に関する主要な情報は，人物の名前，職業，生年月日などの人間にとっての基本的な情報と，その人物の職業に固有な情報に大きく分けられる．例えば，職業が政治家ならば，どの政党に属するか，どの議会の議員なのかは，非常に重要な情報となる．我々は，WWW上に存在する職業別人名リストを利用してこれらの情報を自動収集する方法を提案する．

氏名	生年月日	所属党派	住所	電話
伊東景治	S27.02.24	無所属	古御堂142	54-2736
橋本文一	S25.10.22	日本共産党	若葉3763	54-1887
金屋栄次	S19.0515	無所属	生地吉田9659	56-8835
辻　泰久	S22.08.28	無所属	山田720-1	54-0248

図１　　表形式の職業別人名リストの例

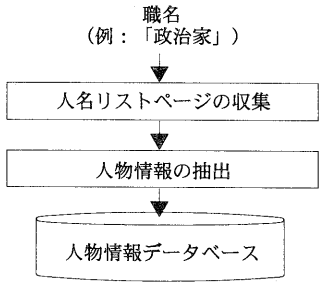


図２　　システム構成

2.1 職業別人名リストの特徴

特定の職業に属する人物の主要情報を簡潔に表現したリストを職業別人名リストと呼ぶ．WWW上には，職業別人名リストが多数存在し，その多くは，表形式**で記述されている．例を図１に示す．本システムでは，このような表形式の職業別人名リストを情報源として利用する．

2.2 システム構成

システムの構成を図２に示す．本システムは，次に示す２つのモジュールから構成される．

(1) 人名リストページの収集
「政治家」などの職名を入力としてWWWから人名リストがあるページ（人名リストページ）を収集する．

(2) 人物情報の抽出
収集したページ内に存在する表形式の人名リストを解析し，人物情報を抽出する．

以下では，２章で人名リストページの収集について述べ，３章で人物情報の抽出について述べる．

3. 人名リストページの収集

人名リストページの収集は，まず，候補ページを収集し，その中に人名リストがあるかどうかを判定することによって行う．候補ページの収集には，検索エンジンとリンク情報を用いる．

---

* Yahoo!のホームページ：http://www.yahoo.co.jp/

** 表形式のリストは，テーブルタグ（<table>，</table>）によって記述される．

表1　職名を表す言葉

職名を表す言葉	例（職名が「政治家」の場合）
1. 職種	「政治家」,「議員」など
2. 職業と関係が深い言葉	「衆議院」,「議会」など

3.1 検索エンジンを用いた候補ページの収集

職名に対して定義されているクエリを検索エンジンに入力することによって,候補ページを収集する．例えば，職名が「政治家」の場合は，クエリとして「議員名簿 or 議員一覧 or 議員紹介」を用いる．

3.2 リンク情報を利用した候補ページの収集

検索エンジンを用いて収集したページには，ページ内には人名リストを含まず，そのページからリンクされているページに人名リストを含むものがある．例えば，収集したページ内に存在する文字列「議員名簿」が他のページへのハイパーリンクをもっており，そのリンク先ページに議員名簿（人名リスト）が存在する場合などがある．このようなリンク先ページからも人名リストを収集するために，リンク情報を利用して，さらに人名リストページの候補を収集する．ここで，リンク情報とは，アンカ *** とそれに併記されたリンク先ページに関する説明文のことをいう．

手順は,(1)から(3)の３ステップからなる（図３）．

(1) 候補ページから次の方法でリンク情報を抽出する．
- リストの１項目にアンカが１つだけ含まれる場合，その項目をリンク情報とする．
- テーブルの行またはセルにアンカが１つだけ含まれる場合，その行またはセルをリンク情報とする．
- HTML ソースの１行にアンカが１つだけ含まれる場合，その行をリンク情報とする．

いずれにも該当しない場合は，アンカのみをリンク情報とする．

(2) 抽出したリンク情報が，次に示す(a)から(c)の条件のいずれかを満たす場合，そのリンク先のページを候補として収集する．

(a) アンカの末尾が，職名を表す言葉（表１）である．（例：「県議会」,「県議員」）

(b) リンク情報が，職業と関係が深い言葉（表１参照）を含み，その後にホームページを表す言葉（「HomePage」など9語）を含む.(例：「衆議院のホームページ」)

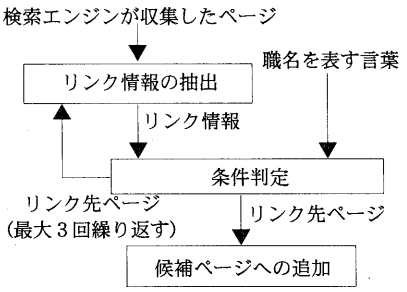


図３　リンク情報を利用した候補ページの収集手順

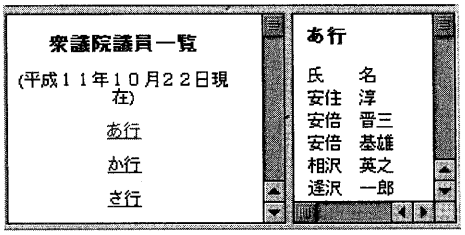


図４　50音で分類された人名リストの例

(c) リンク情報が，職種である言葉（表１参照）を含み，その後に次の言葉のいずれかを含む．
- リストを表す言葉（「名簿」など９語）
- 検索サービスやリンク集を表す言葉（「検索」「リンク」など４語）
- 先の２つの言葉を英語やローマ字にしたもの（URL と照合する）

(3) 新たに候補ページが得られた場合は，そのページに対して,(1)と(2)の処理を同様に行う．これは，最大３リンク先まで行う．

3.3 人名リストの有無判定

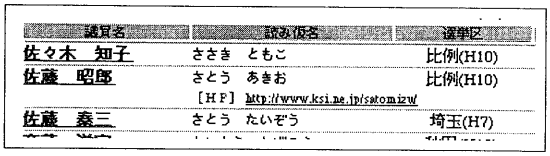
検索エンジンやリンク情報を利用して収集した候補ページには，必ずしも人名リストが存在するとは限らない．そこで，それらから人名リストがあるページのみを選別する．

しばしば，人名リストは50音で分類されることがある．例を図４に示す．このことを考慮し，次の(1),(2)の方法で人名リストの有無判定を行う．

(1) 候補ページを調べ,次の条件のいずれかを満たす場合，候補ページに人名リストがあるとする．

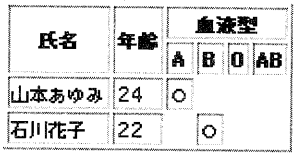
(a) ページ内の見出し，ページタイトル，逆リンク情報のいずれかに,職名を表す言葉とリストを表す言葉が含まれる．

*** アンカ：アンカタグ(<a>，)の範囲の文字列を指す．（ アンカ ）



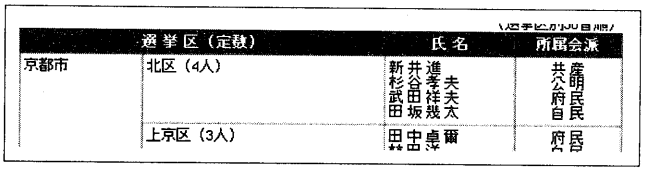
議員名	読み仮名	選挙区
佐々木　知子	ささき　ともこ	比例(H10)
佐藤　昭郎	さとう　あきお	比例(H10)
	[HP] http://www.ksi.ne.jp/satomizu/	
佐藤　森三	さとう　たいぞう	埼玉(H7)

図5　明記されたどのフィールド名にも属さないデータ（「[HP] http://www.ksi.ne.jp/satomizu/」）がある表



氏名	年齢	血液型			
		A	B	O	AB
山本あゆみ	24	○			
石川花子	22			○	

図7　1つのフィールド名（「血液型A」など）が複数のセルからなる表



選挙区（定数）		氏名	所属会派
京都市	北区（4人）	新井進 杉谷孝夫 武田祥夫 坂井幾太	公明 自民 自民 民主
	上京区（3人）	田中卓雄	自民

図6　tdタグでなく改行を用いた表

氏名	年齢	血液型	血液型	血液型	血液型
氏名	年齢	A	B	O	AB
山本あゆみ	24	○			
石川花子	22		○		

図8　図6の表を標準化した表

(b) そのページが，50音で分類されている．（より正確には，同一ページへリンクしている50音アンカ*があるか，アンカでない50音文字列がある．）

(2) 候補ページ内に50音アンカが存在し，次の条件のいずれかを満たす場合，そのリンク先ページに人名リストがあるとする．

・50音アンカの一段上位の見出し**が職名を表す言葉を含む．

・候補ページが(1)の(a)の条件を満たす．

4. 表解析による人物情報の抽出

前章の方法で収集した人名リストページに存在する表形式の職業別人名リストを解析し，得られた人物情報をデータベースに格納する．ここで，解析対象とする表形式の人名リストは，テーブルタグを用い，フィールド名が明記されたものである．解析手順は，まず，表から人物情報を抽出する．次に，表の見出しから得られる人物情報を獲得する．最後に，これらの結果をまとめ，データベースに格納する．

4.1 表からの人物情報の抽出

WWW上には，様々なレイアウトの表がある．例を図5から図7に示す．人物情報の抽出では，このような様々な表の構造を正しく把握する必要がある．以下に，人物情報の抽出手順を示す．

(1) 表とその見出しの抽出

テーブルタグは，表としてではなく，レイアウトに使用される場合もある．ここでは，単純な方法で表の抽出を行う．まず，罫線のある表を抽出する．次に，内部にテーブルタグを含まない表を抽出する．表の見出しは，抽出した表の上部からHTMLタグを利用して抽出する．

(2) 表の標準化

図1のような表を標準形式の表とし，次の(a)から(d)の手順で，表のレイアウトを標準化する．その際，後の処理で必要となる表の方向やフィールド名の範囲も調べる．

(a) 1セルが複数のセルを越えていない表にする．具体的には，1行または1列に1セルしかない場合は，そのセルを削除する．複数のセルにまたがるセルが存在する場合は，そのセルを分割して同じ値をもつ複数のセルを設定する．例えば，図7の表は図8の表に標準化する．

(b) 表の方向を調べる．

人物に関する属性名を用いて，縦，横のどちらの方向であるかを決める．人物に関する属性名として，「氏名」，「年齢」などの基本的な属性名と，職名が「政治家」ならば「選挙区」，「会派」などの職業固有の属性名の両者を用いる．

(c) フィールド名の範囲を調べる．

横方向の表の場合は，1行目と同じ値が何行まで続いているかを調べ，その最大値をフィールド名の範囲とする．縦方向の表の場合は，横方向の場合の行を列として同様の処理を行う．

(d) 表が結合している場合は，1つの表にする．

---

* 50音アンカ（50音文字列）：「あいうえお順」，「50音順」，「あ行」，「あ」，「あ～お」のような言葉のこと．

** 一段上位の見出し：図4の「衆議院議員一覧」のように，何の50音であるかを示す言葉のこと．

広島県議会議員（定数70人）

※氏名をクリックするとプロフィールがご覧になれます。

選挙区	会派	氏名	郵便番号	住所	電話番号
広島市中区	自民	林（はやし） 正夫（まさお）	730-0052	中区千田町三丁目6-32	082-244-0884
	公明党・県民会議	中田（なかた） 選（せん）	730-0847	中区舟入南四丁目17-9	082-295-7744

```
<林正夫>
  <選挙区><データ>広島市 中区</データ></選挙区>
  <会派><データ>自民</データ></会派>
  <氏名><データ
        リンク先ページ={"http://www.hiroshima-cdas.or.jp/pref/gikai/giin/giinprof/hayasi.html"}
        画像ファイル={"http://www.hiroshima-cdas.or.jp/pref/gikai/gif/hayasi.gif"}
      >林 正夫</データ></氏名>
  <郵便番号><データ>730-0052</データ></郵便番号>
  <住所><データ>中区千田町三丁目6−32</データ></住所>
  <電話番号><データ>082-244-0884</データ></電話番号>
  <職業>政治家・広島県議員</職業>
  <出典><データ URL="http://www.hiroshimacdas.or.jp/pref/gikai/giin/giin_mei.html">
        議会とは</データ></出典>
</林正夫>
```

図9　上の表から抽出した人物情報

(3) 標準形式の表からの人物情報の抽出

人名別に，フィールド名とそれに対応する値を対にしたものを人物情報として抽出する．ここでは，横方向の表の場合について説明する．縦方向の表の場合は，横方向の場合の行を列として同様の処理を行う．

(a) 氏名フィールドがあることを確認する．

フィールド名の範囲内に氏名を表す言葉と一致するものがあれば，氏名フィールドがあるとする．氏名を表す言葉には，「氏名」，「名前」，「しめい」，「なまえ」，「名」，職種である言葉（「政治家」，「議員」）を組み合わせたものを用いる．

(b) 1レコードの範囲を調べる．

1レコードが何行かを調べる．1レコードの範囲は，基本的にフィールド名の範囲と同じ行数である．ただし，図7のように違う場合もある．そこで，このような場合に対応するための処理も行う．図7の場合では，1レコードは1行となり，フィールド名は，「氏名」，「年齢」，「血液型A」，「血液型B」，「血液型O」，「血液型AB」となる．

(c) 1レコードづつ情報を抽出する．

フィールド名と同じ順に値があるという前提で，フィールド名とそれに対応する値を対にし，氏名フィールドの値別に人物情報として収集する．その際，フィールド名に括弧があり，値にも括弧がある場合は，括弧前と括弧内に各々分割し対応させる．また，様々なレイアウトの表に対応するため，いくつかの例外処理も行う．

4.2 表の見出しから得られる人物情報の獲得

職業別に表の見出しから得られる人物情報を獲得する処理を行う．例えば，職名が「政治家」の場合は，表の見出しから衆議院議員や石川県議員などの職名のサブカテゴリをパターンマッチングにより抽出する．表の見出しから獲得できない場合は，ページタイトルや逆リンク情報も利用する．なお，職名が「政治家」の場合で，それでもサブカテゴリが獲得できない場合は，URLから地名が特定できれば，サブカテゴリをその地名の議員とする．

4.3 人物情報データベースへの格納

表から人名別に抽出した人物情報に，職業情報（職名，サブカテゴリ名）と出典情報（表を抽出したページのURLとページタイトル）を加えて，データベースへ格納する．格納するデータの例を図9に示す．ただし，不適切なデータが格納されないように，次の2つの条件を満たす場合のみ格納することにする．

- 表の見出し，ページタイトル，逆リンク情報のいずれにも「候補」または「予定」を含まない．
- 表の見出し，ページタイトル，逆リンク情報のいずれかに職業のサブカテゴリ名か職業と関係が深い言葉を含む．または，フィールド名に職業固有の属性名がある．

表2　サブカテゴリ別収集状況

職業	収集数	職業のサブカテゴリ総数	議員定数との比較結果
国会議員	2	2	すべて不一致
都道府県議員	13	47 (人手 18)	すべて一致
市町村区議員	22	3380	すべて一致

表3　国会議員定数との比較

サブカテゴリ名	収集人数	定数
衆議院議員	546	500
参議院議員	254	252

5. 実験と検討

「政治家」と「著述家」を対象として，人物情報を自動収集する実験を行った．なお，前節で述べた表解析は都道府県議員リストを参考にして作成したものである．

5.1 「政治家」である人物の情報収集

国会議員と地方議員の情報を収集する実験を行った．「政治家」用のデータとして以下のものを用いた．

(1) 検索エンジンへのクエリ：「議員名簿 or 議員一覧 or 議員紹介」

(2) 職名を表す言葉
- ・職種である言葉：「政治家」，「議員」，「seijika」，「giin」の4語
- ・職業と関係が深い言葉：「衆議院」，「議会」，「shugi」など6語

(3) 職業固有の属性名：「政党」，「選挙区」など4語

また，表の見出しから獲得する人物情報は，衆議院議員，石川県議員などの職業のサブカテゴリとする．

収集した人物情報をサブカテゴリ別に整理した結果を表2に示す．国会議員のみ，議員定数と一致していなかった．表3に国会議員の定数との比較結果を示す．いずれの場合も定員以上の人数が見つかっている．これは，字体の違いや旧リストと新リストの混合による．都道府県議員は，サブカテゴリ総数が47であるが，本システムで利用した検索エンジンを用いて人手で探したところ，そのうち18に対して議員リストがあることが確認された．システムはそのうち13を収集した．市町村区議員は，サブカテゴリ総数が3380で，そのうち21を収集した．

- ・受賞名（6）
  - ・野間文芸賞（1）
  - ・山本周五郎賞（1）
  - ・日本推理サスペンス大賞（1）
  - ・吉川英治文学新人賞（1）
  - ・サントリーミステリー大賞（1）
  - ・鮎川哲也賞（1）
- ・ミステリー作家（1）
- ・作家（1）

（）内の数字：ページ数

図10　賞またはサブカテゴリ別収集状況

5.2 「著述家」である人物の情報収集

受賞作品がある著述家の情報を収集する実験を行った．「著述家」用のデータとして以下のものを用いた．

(1) 検索エンジンへのクエリ：「(一覧 or リスト) and 著者 and 賞」

(2) 職名を表す言葉
- ・職種である言葉：「作家」，「著者」，「受賞作家」，「受賞者」の4語
- ・職業と関係が深い言葉：「文学」，「ミステリー」，「賞」など5語

(3) 職業固有の属性名：「作品名」など8語

また，表の見出しから獲得する人物情報は，主に受賞名とする．

収集した人物情報を賞または職業のサブカテゴリ別に整理した結果を図10に示す．6種類の受賞者リストから人物情報を収集したが，フィールド名に対する値に不適切なものが存在した．不適切の原因は2つある．1つは，著者フィールドにある値を氏名としたため，編集局名などが氏名となった．もう1つは，フィールド名と値を対応させる際に，フィールド名と同じ順に値があるという前提で行うため，図11のような表を解析した場合は，不適切な対応が得られた．

5.3 検討

「政治家」に対しては，人物情報の収集状況はよかった．この理由として，議員リストは表形式のものが多いことと，リンク情報を利用して議員リストを探す場合，「議会ホームページ」，「議員名簿」，「あ行」のような言葉が手がかりとして有効に働くことが挙げられる．

一方，「著述家」に対しては，それほど多くの人物を収集できなかった．これは，受賞者リストが賞名で分類されることが多いことと，リストの見出しが賞名のみでリストを表す言葉をあまり用いないことによる．より多くのリストを収集するためには，次

に示す改良が必要である．

・50音以外で分類されたリストも収集する．

・人名リストの有無判定で，手がかり語（職名を表す言葉とリストを表す言葉）の有無だけでなく，表形式のリストの有無も調べる．

・賞名をシステムに与える．

・人名リストがあるページをリンクしているページを調べる．

回	年度	受賞名	受賞作品	著者	出版社	出版年
		選考委員				
第1回	昭和58年(1983年)	大賞	虹へ、アヴァンチュール	愛羽十九哉	文藝春秋社	83/06
		読者賞	桜子は帰って来たか	麗羅	文藝春秋社	83/06
		佳作賞	二度のお別れ	島川博行	文藝春秋社	84/09
		阿川弘之、開高健、小松左京、田辺聖子、都筑道夫				
		大賞	運命交響曲	由良三郎	文藝春秋社	84/06

図11　表解析で失敗した表

6. 人名からの人物プロフィールの収集

職業別人名リストには，表形式のもの以外に，人物プロフィールを列挙したものがある．例を図12に示す．このような列挙形式の職業別人名リストを情報源とし，フォーマット情報を利用して，人物プロフィールを収集するシステムを作成した．システムの概要を図13に示す．ここでは，ページ内にある入力人物のプロフィールを抽出する方法と，38名に対して行った実験について述べる．

6.1 人物プロフィールの抽出

人名を入力として，ページ内にある入力人物のプロフィールを次の(1)から(5)の手順で抽出する．

(1) テーブルタグが次に示す(a)または(b)のために使用されている場合，別のタグに置換する．

(a) 見出しやナビゲーションバー

(b) ページの割り付け（具体的には，テーブルタグの範囲に，(a)に該当するテーブルタグ，HTMLのリストや見出しを示すタグ，「です.」などの文末表現の言葉のいずれかがあるものとする.）

(2) HTMLソースにフォーマット情報を示す4種類のタグ（空行，改行，罫線，インデント）を挿入する．

(3) 次に示す見出しの3つの特徴を用いて，見出しの部分をマークアップする．

・文字列の特徴：短く，記号から始まる．

・HTMLタグの特徴：文字を強調するタグ(見出しタグや太字タグなど)を用いる．

・位置の特徴：インデントの前や罫線の後にある．

(4) 見出しに入力名があれば，その本文を抽出する．本文の領域は，入力名がある見出しの次の行からこの見出しがもつ特徴（(3)で示した3つの特徴）と同じ特徴をもつ見出しまたは上位の見出しまでとする．もし，この方法で本文となる領域がないと判定された場合は，見出しと判定した入力名がある文字列のみに対して(5)の処理を行う．

(5) (4)で抽出した部分に「本名」，「生(ま)れ」，「出身」のいずれかがあればプロフィールであると判定する．

長谷川如是閑（にょぜかん）

本名・万次郎。ジャーナリスト。

長谷川龍生

本名・名谷竜夫。

長谷　健

本名・藤田正俊。

図12　列挙形式の職業別人名リストの例

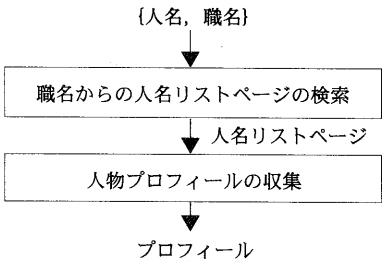


図13　システム構成

6.2 実験

前章の「著述家」を対象とした実験で収集した人名からランダムに選んだ38名を用いて，これらの人物のプロフィールを収集する実験を行った．プロフィールの抽出対象となるページは，「著述家」で収集した人名リストページと，空白のない人名を入力としたgooの検索結果のページ（最大件数100件）である．

収集した人物プロフィールの例を図14に示す．また，実験結果を表4に示す．全体として，38名のうち22名に対してプロフィールが収集された．また，職業別人名リストに対してシステムが収集した16名のプロフィール(22件)を調べたところ，プロフィールとして不適切なものが3件存在した．不適切なものの例を図15に示す．その原因は，キーワードの有無のみでプロフィールを判定したことにある．

[見出し] 折原一（おりはらいち）
[本文(428)]一九五一年一一月一六日埼玉県久喜市生まれ。早稲田大学第一文学部卒業。JTBに入社し、旅行雑誌「旅」などの編集に携わる。退社後、八八年、「五つの棺」でデビュー。〈作品〉『倒錯の死角』88・東京創元社刊。『倒錯のロンド』89・講談社刊。『灰色の仮面』90・講談社刊。『異人たちの館』93・新潮社刊。『沈黙の教室』94・早川書房刊＝第四十八回日本推理作家協会賞（長編部門）受賞。『漂流者』96・角川書店刊。『遭難者』97・実業之日本社刊、など。

図14　抽出した人物プロフィールの例

[見出し] [img alt="記者1" src="./kisya-kao/kokubu.gif"]「霞町物語」（浅田次郎、講談社）
「ストリッパー」（二代目一条さゆり、幻冬舎アウトロー文庫）
[本文(974)]・長～い無読書状態の中からようやっと抜け出して読んだのが、日本から来

（省略）

と、甘酸っぱい思い出が広がる。福岡市出身の二代目がつづる時代は「その後」のことで良くは知らないが、旅から旅の地方巡業、香港映画を追っての香港通い…などなど、興味深く読ませてもらった。二代目は昨年秋から広州の大学に留学中とか。取材でも何でも良いから一度是非お目にかかりたいものだ。

図15　抽出したプロフィールが不適切である例（入力名「浅田次郎」の場合）

表4　人物プロフィールの収集状況

対象ページ	職業別人名リスト	gooの検索結果
プロフィール収集人数 / 入力名総数	16 / 38	13 / 38

7. おわりに

本稿では，WWWから人物情報を自動収集する方法として，職業別人名リストを利用した２つの方法を提案した．１つは，表形式の職業別人名リストから表解析を行って人物情報を収集する方法である．もう１つは，列挙形式の職業別人名リストからフォーマット情報を利用して人物プロフィールを収集する方法である．これらのシステムは，職業別人名リストを情報源とするため，基本的な人物情報と職業固有の人物情報を同時に収集できるという利点がある．しかし，収集できる人物に限りがある．今後は，WWW上の個人ページを対象とした人物プロフィールの収集を行う予定である．

既存の人物情報を提供する検索サービスとして，個人のホームページを探し出すAhoy![2]がある．これに対して，提案手法は，ページを収集するだけでなく，人物情報の抽出も行う点に違いがある．

ウェブページから情報を抽出する場合，HTMLタグを手がかりとして用いることができる[3]．しかしながら，HTMLタグが不適切に使用されている場合も多く，完全に信用することはできない．情報抽出などの応用処理を容易にするために，あるタグセットを設定し，それに従い，あらかじめページにタグを付加しておく方法も提案されている[4][5]．しかし，我々は，そのようなタグを人手で付加することは現実的ではないと考え，そのようなタグを仮定せずに自動抽出を実現する立場をとっている．

テキストから人物情報を抽出する研究としては，新聞記事から表層パターンに基づいて人物情報を抽出する西野らの研究[6]がある．我々は，WWW上の職業別人名リストを利用するという点に大きな違いがある．

謝辞

本研究の一部は，科学技術振興事業団からの受託研究「利用目的に応じた情報の組織化と自動編集」の助成によるものである．

参考文献

[1] 山本あゆみ，佐藤理史．WWW上の職業別人名リストを利用した人名の収集．情報処理学会第59回全国大会，Vol.3, pp.119-120, 1999.

[2] Jonathan Shakes, Marc Langheinrich and Oren Etzioni. Dynamic Reference Sifting: a Case Study in the Homepage Domain. WWW6, pp.189-200, 1997.

[3] Dan DiPasquo. Using HTML Formatting to Aid in Natural Language Processing on the World Wide Web. enior Honors Thesis, School of Computer Science, CMU, 1998.

[4] 渡辺日出雄．Web文書に対する言語処理の問題点と言語処理を援助するタグセットについて．情処研報，NL127-13, pp.95-100, 1998.

[5] Dayne Freitag. Information Extraction From HTML: Application of a General Learning Approach. Proceedings of the 15th National Conference on Artificial Intelligence, AAAI, 1998.

[6] 西野文人，落谷亮．新聞記事からの人物・企業情報の抽出．情処研報，NL127-17, pp.125-132, 1998.

盛光印刷所