

Title	ワールドワイドウェブにおける人物検索の実現
Author(s)	山本, あゆみ
Citation	
Issue Date	2000-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1322
Rights	
Description	Supervisor:佐藤 理史, 情報科学研究科, 修士

修 士 論 文

ワールドワイドウェブにおける人物検索の実現

指導教官 佐藤 理史 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

810121 山本 あゆみ

2000 年 2 月 15 日

要 旨

ワールドワイドウェブ上に存在する人物情報を検索するシステムを実現するために、2つの方法を提案する。第1の方法は、ある職業の人名録を情報源として表解析により人物情報データベースを自動生成する方法である。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、その職業の人名録を収集する。次に、収集された人名録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、与えられた人名からその人物を紹介したテキストが存在するページを収集し、それらのページからレイアウト情報を利用してその人物を紹介した短いテキストを抽出する方法である。

目次

1	序論	1
2	人物情報と人物検索システム	3
2.1	人物情報	3
2.2	人名録	4
2.3	ウェブ上の人物情報と人名録	4
2.4	人物検索システムの設計	8
2.5	システム構成	10
3	職名からの人物情報の収集	15
3.1	人物情報収集の概要	15
3.2	人名録ページの収集	16
3.2.1	検索エンジンを用いた候補ページの収集	16
3.2.2	リンク情報を利用した候補ページの収集	17
3.2.3	人名録の有無判定	20
3.3	表解析による人物情報の抽出	24
3.3.1	表からの人物情報の抽出	24
3.3.2	表の見出しからの人物情報の抽出	32
3.3.3	人物情報データベースへの格納	34
3.4	実験と検討	34
3.4.1	実験 1 : 職名が「政治家」の場合	34
3.4.2	実験 2 : 職名が「著述家」の場合	39
3.4.3	検討	41

4	人名からの人物情報の収集	42
4.1	人物情報収集の概要	42
4.2	候補ページの収集	43
4.3	レイアウト解析による人物情報の抽出	44
4.3.1	前処理	45
4.3.2	人物紹介ページの判定	46
4.3.3	テキストからの人物情報の抽出	48
4.3.4	個人情報リンクの抽出	52
4.3.5	結果作成	52
4.4	実験と検討	52
4.4.1	実験 1：人名と職業を指定した場合	56
4.4.2	実験 2：人名のみを指定した場合	58
4.4.3	検討	61
5	検討と関連研究	62
5.1	検討	62
5.2	関連研究	63
6	結論	65
A	実験結果	71

第 1 章

序論

ワールドワイドウェブ上では、誰もがいつでも情報を発信でき、個人や企業などから様々な情報が提供されている。これらの情報は、分散化されていて、その量は膨大である。このような情報源から求める情報を探し出すために、ロボット型検索エンジンと呼ばれる汎用の検索システムが開発されてきた。この検索システムは、「ロボット」または「スパイダー」と呼ばれるプログラムによって、大量のウェブページを自動収集し、それらページに対するキーワード検索を提供する。そのため、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。したがって、利用者が効率良く情報収集するのは、困難である。

この問題を解決する 1 つの方法として、検索対象のカテゴリを限定した検索システムが考えられる。検索対象の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検索システムを実現することを目的とする。このシステムを実現するために、人物の主要情報を自動収集する次の 2 つの方法を提案する。

(1) 職業別人名録を利用した情報収集

ウェブ上には、ある職業に属する人物の主要情報をまとめて整理した職業別人名録が多数存在し、その多くは、表形式で記述されている。このような職業別人名録は、

人名がわからなくても、職名がわかれば、収集可能である。そこで、職名からその職業の人名録を見つけ出し表解析を適用することで、効率的に多くの人物に関する主要情報を収集する方法を提案する。この方法は、オフラインで実行できるため、あらかじめ収集した情報をデータベース化し、検索時に、このデータベースを参照して利用者に情報提供するアプローチをとることができる。

(2) (1)を補完する形で人名からの情報収集

人名録を利用する方法は、効率良く人物情報を収集することができるが、人名録に掲載されている人物の情報しか収集できないという問題がある。そこで、人名からその人物の主要情報を収集するもう1つの方法を提案する。この方法では、人物の主要情報がテキストで記述されているページを情報源とし、テキスト形式の人物情報を収集する。テキスト形式の人物情報は、人名の見出しと、その人物を紹介したテキストからなることが多く、そのレイアウトにいくつかのパターンが見られる。このレイアウトパターンを利用して情報抽出を行う。この方法による収集は、人名が与えられないと実行できないため、オンラインで実行する。

本論文の構成は、以下の通りである。まず、第2章では、検索の対象とする人物情報について述べ、この人物情報を検索するシステムについて述べる。第3章では、職名から人物情報を自動収集する方法と、この方法を用いて行った実験の結果について述べ、この方法の有効性を検討する。第4章では、人名から人物情報を自動収集する方法と、この方法を用いて行った実験の結果について述べ、この方法の有効性を検討する。第5章では、人物検索システムの有効性について検討し、関連研究について述べる。最後に、第6章で、結論を述べる。

第 2 章

人物情報と人物検索システム

本章では , まず , 本研究で検索の対象とする人物情報とはどのような情報であるかを明確にし , それがウェブ上にどのような形式で掲載されているかについて述べる . 次に , そのような人物情報を検索するシステムの概要に述べる .

2.1 人物情報

いま , ある論文を手に行っているとしよう . このとき , その論文の著者がどのような人間であるか知りたいと思うことは , 比較的よくあることである . このような要求に答えるために , 多くの論文には , **著者紹介**が存在する . ここには , 著者がどのような人物であるかが簡潔に書かれている . 本論文では , 著者紹介に書かれるような , その人物に関する主要な情報のことを**人物情報**と呼ぶ .

人物情報は , 次に示すように , 情報内容 , および , 記述形式 , の 2 つの軸によって分類することができる .

(1) 情報内容による分類

(a) 多くの人間に共通した属性

名前 , 職業 , 生年月日 , 現住所など .

(b) それぞれの職業に固有な属性

例えば , 著述家ならば主要著書 , 政治家ならば所属議会 , 選挙区など .

(2) 記述形式による分類

(a) テキスト形式

人物情報をテキストとして記述する．著者紹介などが典型的な例である．

(b) 属性 - 属性値集合

人物情報を属性-属性値の集合として記述する．表形式で書かれる場合が多い．

本研究では，このような人物情報をウェブから収集し，利用者に提供することを目的とする．

2.2 人名録

人物情報には，論文の著者紹介のように，単独で存在するもの以外に，多くの人物に関する情報をまとめて整理した形態で存在するものがある．これらは，通常，人名録，あるいは，名簿と呼ばれることが多い．典型例として，紳士録や各種の人名辞典，職員録などがある．

人名録には，ある職業に属する人物の情報を集めた**職業別人名録**が存在する．例えば，『現代政治家人名辞典』は，1979年以降に実施された公選で選ばれた政治家3671人を収録した人名辞典である．この人名辞典には，肩書き・職業（所属政党・選挙区），専攻分野，生年月日，没年月日，出生（出身）地などの情報が記載されている．このうち，肩書き・職業（所属政党・選挙区）は，2.1節で述べた職業固有属性の典型例である．このように，職業別人名録では，職業固有属性が一意に定まるため，人物情報として掲載すべき情報を標準化しやすいという長所をもつ．

2.3 ウェブ上の人物情報と人名録

近年，多くの人物情報がウェブ上に掲載されるようになってきている．これらの情報は，次に示す3つの形態で存在する．

(1) 個人のホームページ

個人のホームページには，本人の略歴が掲載されていたり，自己紹介のページ（図2.1参照）など本人の情報を掲載したページへのリンクが存在したりすることが多い．

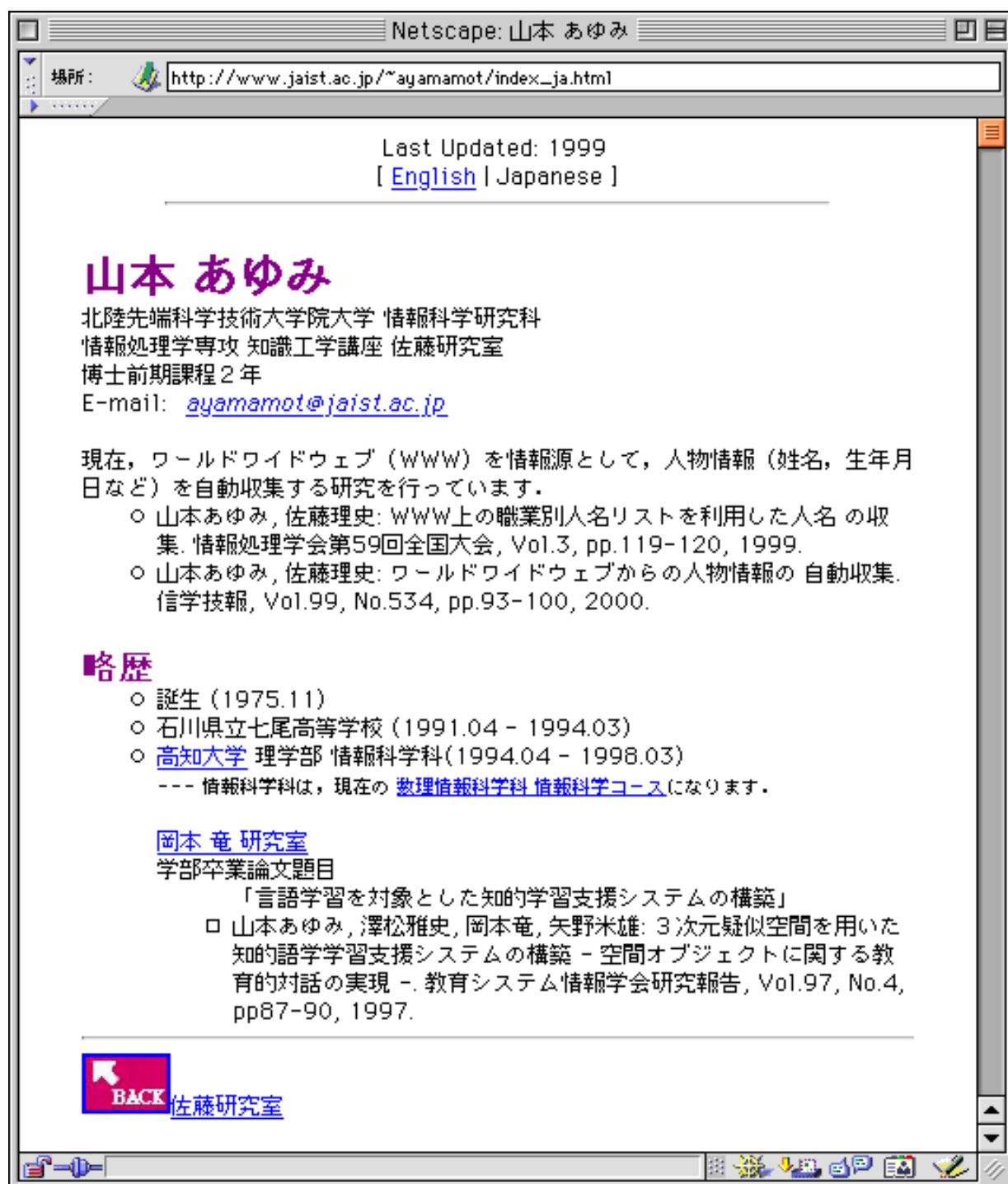


図 2.1 自己紹介のページの例



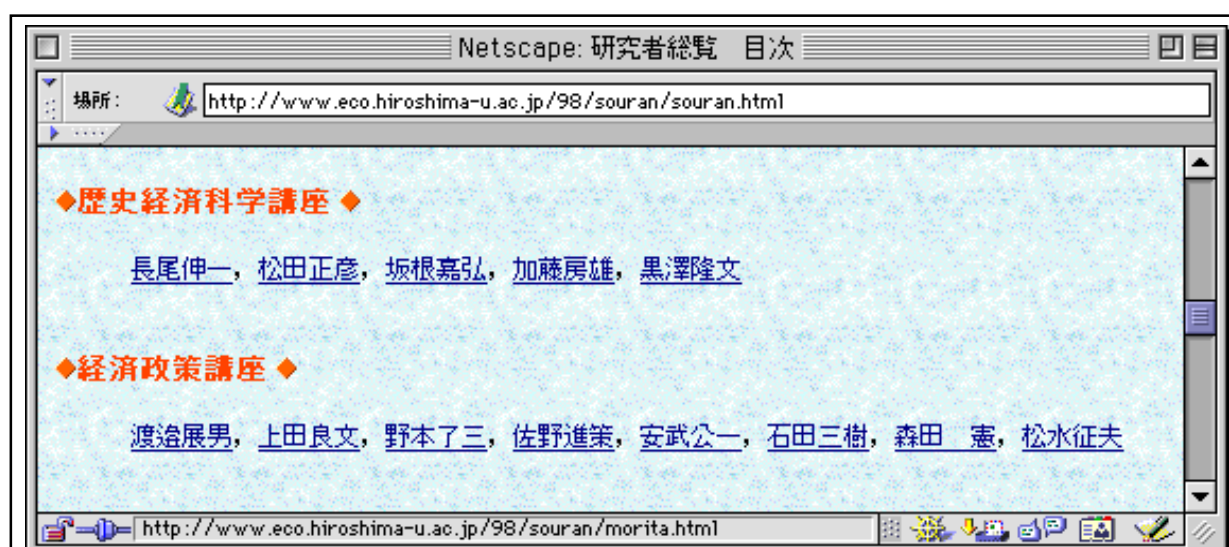
図 2.2 表形式の人名録が存在するページの例

(2) 人名録ページ

人名録が存在するページを人名録ページと呼ぶ。人名録は、1 ページ、または、複数ページで構成されている。複数ページから構成される場合は、人名録を 50 音などで分割する場合と、それぞれの人物に 1 ページを充てる場合がある。例を図 2.2、図 2.3 示す。図 2.2 が示すページは、自由民主党所属衆議院議員名簿一覧（人名録）の一部であり、氏名があ行から始まる名簿が掲載されている。上部にある「衆議院議員 - あ行 - か行 - さ行 ...」の例えば「か行」をクリックすると、氏名がか行から始まる議員名簿がブラウザに表示される。図 2.3 の上部に示すページは、広島大学経済学部研究者総覧（人名録）であり、各講座の研究者の氏名が掲載されている。例えば、その 1 つである「長尾伸一」をクリックすると、その人物の主要情報が掲載されたページ（図 2.3 の下部に示すページ）が表示される。

(3) 人名録データベース

人名録は、データベースとして存在する。利用者が入力した人名などのキーワードからデータベース検索を行い、その結果（人物情報）を提供する。例えば、静岡県



上に示すページの「長尾伸一」をクリックすると下に示すページが表示される



図 2.3 複数ページで構成される人名録の例

科学技術振興財団のサイトにある研究者データベース¹などがある。

これらの形態において、人物情報のレイアウトには、次に示すいくつかのパターンが見られた。

(1) 1名の人物情報によく用いられるレイアウト

- 属性名とその値の集合からなり、表、または、属性名とその値の間に「：」などの区切りとなる文字を記述したものの集合（図2.4参照）、で表現される。
- 人名の見出しと、その人物を紹介したテキスト（図2.5参照）からなり、次に示す形で表現される。
 - ・ 人名の見出しがあり、その次の行から人物の紹介を記述する。
 - ・ 人名の見出しの前後に、空行、または、罫線、があり、その次の行から人物の紹介を記述する。
 - ・ 人名の見出しの後に、「：」などの区切り文字があり、続けて人物の紹介を記述する。
 - ・ 行をインデントして人物の紹介を記述する。
- 1ページに1名の人物情報が記述されている場合は、ページの始めに人名の見出しが存在する。

(2) 複数名の人物情報によく用いられるレイアウト

一覧表（図2.2参照）、または、1名の人物情報の列挙、で表現される。後者の場合は、人物情報の間に、空行、または、罫線、があることが多い。

2.4 人物検索システムの設計

以上で述べたようなウェブ上での人物情報の特徴を考慮し、人物検索システムを実現するための人物情報収集方法として、次の2つの方法を採用する。

¹ <http://www.sstf.or.jp/researcher/>



図 2.4 属性 - 属性値集合の人物情報の例

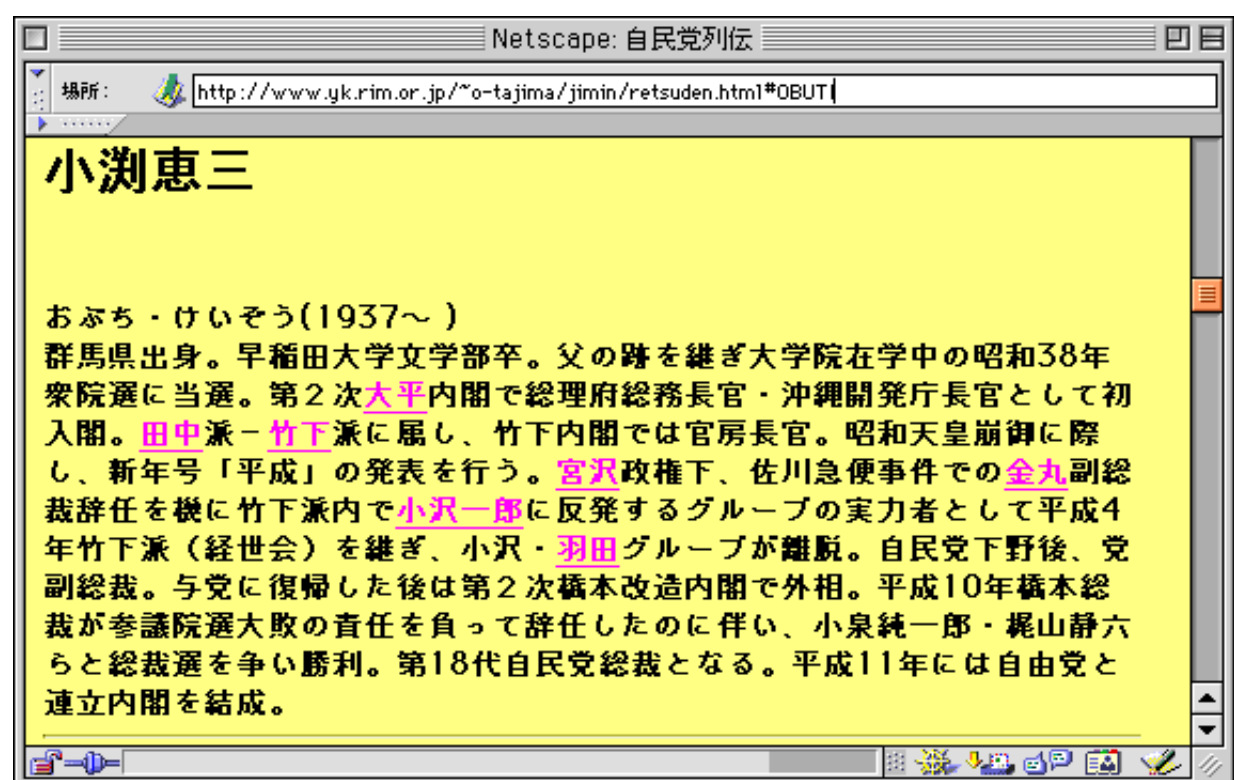


図 2.5 テキスト形式の人物情報の例

(1) 職業別人名録を利用した情報収集

ウェブ上では、職業別人名録が多数存在し、その多くは、表形式で記述されている(図2.2参照)。したがって、人名録に表解析を適用することで、多数の人物の主要情報を効率良く収集できる。また、職業別人名録は、人名がわからなくても、職名がわかれば、収集可能である。すなわち、職名から人物情報を自動収集することが可能となる。そこで、収集した人物情報をあらかじめデータベースに格納し、検索時に、このデータベースを参照するアプローチをとる。

(2) (1)を補完する形で人名からの情報収集

人名録を利用する方法は、効率良く人物情報を収集することができるが、人名録に掲載されている人物の情報しか収集できないという問題がある。この問題を解決するために、人名からその人物の主要情報がテキストで記述されているページを見つけ出しレイアウト解析を適用することで、その人物の主要情報を収集する方法を提案する。この方法による収集は、人名が与えられないと実行できないため、オンラインで実行する。

2.5 システム構成

本研究で作成したシステムの構成を図2.6に示す。本システムは、次に示す4つのモジュールから構成される。

(1) ユーザインタフェース

検索入力ページを図2.7に示す。検索入力ページには、職名選択肢と、人名入力欄が存在する。人名は必ず入力することとし、職名は指定しなくてもよい。以下に、職名が指定された場合と指定されない場合の動作を示す。

(a) 職名が指定された場合

2段階の出力を行う。まず、人物情報データベースから該当するデータを検索し、その結果をブラウザに表示する。例を図2.8に示す。さらに情報を求める場合は、結果ページの下部にある「さらに調べる場合は、こちら。」の下線部をクリックする。これにより、人名と職名をプロフィール収集モジュールに入力し、このモジュールが収集した人物情報をブラウザに表示する(図2.9参照)。なお、

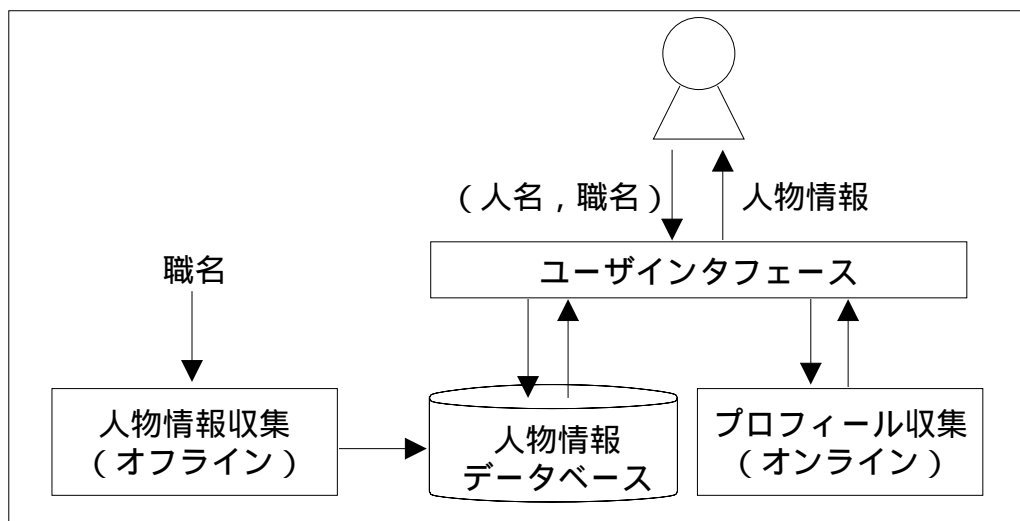


図 2.6 システム構成

データベースに該当データがない場合は，自動的にプロフィール収集モジュールを用いて，その結果をブラウザに表示する．

(b) 職名が指定されない場合

人名をプロフィール収集モジュールに入力し，このモジュールが収集した人物情報をブラウザに表示する．例を図 2.9 に示す．

(2) 人物情報収集

職名を入力とし，まず，その職業の人名録が存在するページ（**職業別人名録ページ**）を収集する．次に，それらのページに存在する表形式の人名録から人物情報を抽出し，人物情報データベースに格納する．人物情報の収集方法については，3章で述べる．

(3) 人物情報データベース

人物情報収集モジュールが収集した属性-属性値の集合で表現された人物情報が格納されている．

(4) プロフィール収集

人名と職名，または，人名のみ，を入力とし，その人物の主要情報がテキストで記述されているページのURLと，それらのページから抽出したテキスト形式の人物情報（**プロフィール**）を出力する．人物情報の収集方法については，4章で述べる．

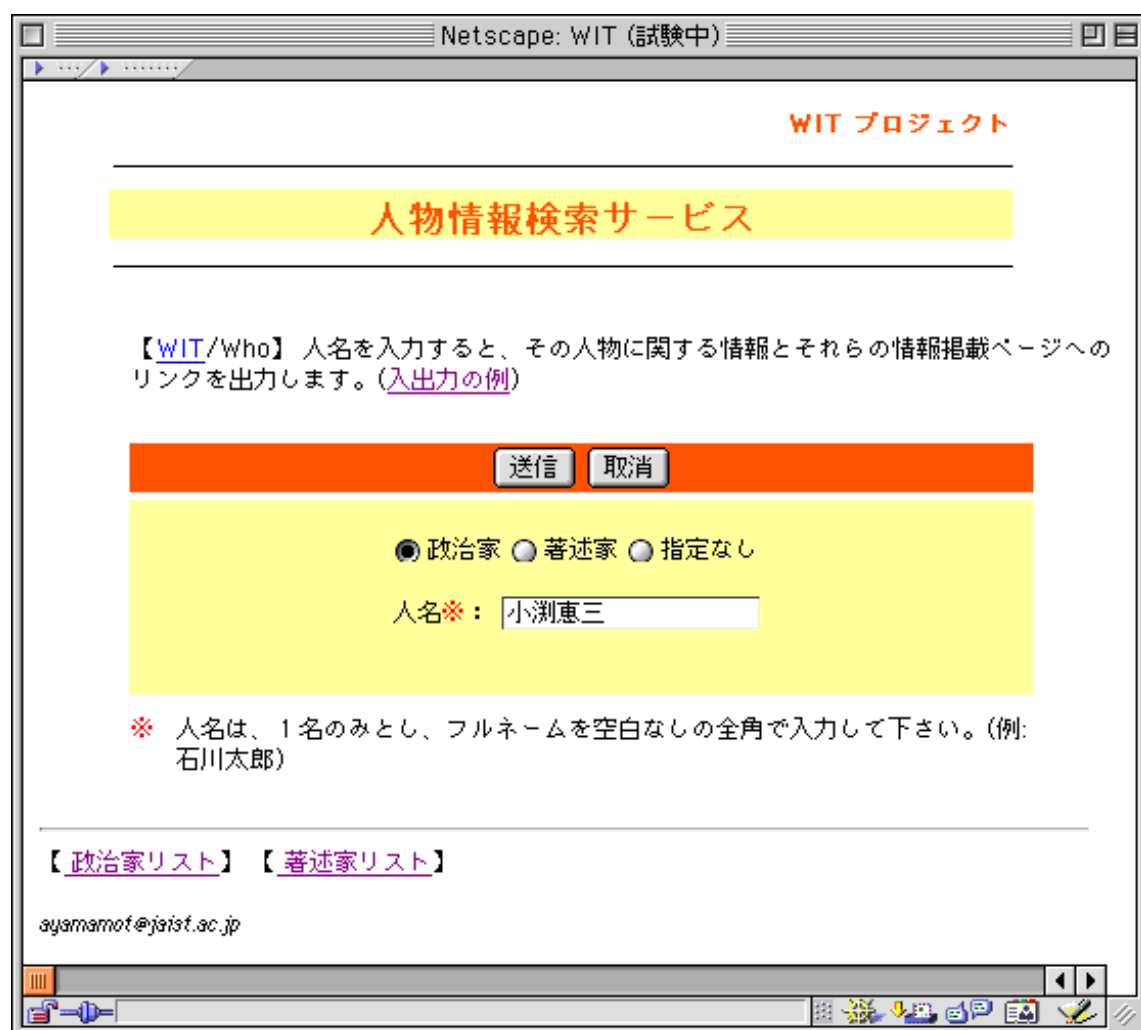


図 2.7 検索入力ページ



図 2.8 データベース検索結果ページ

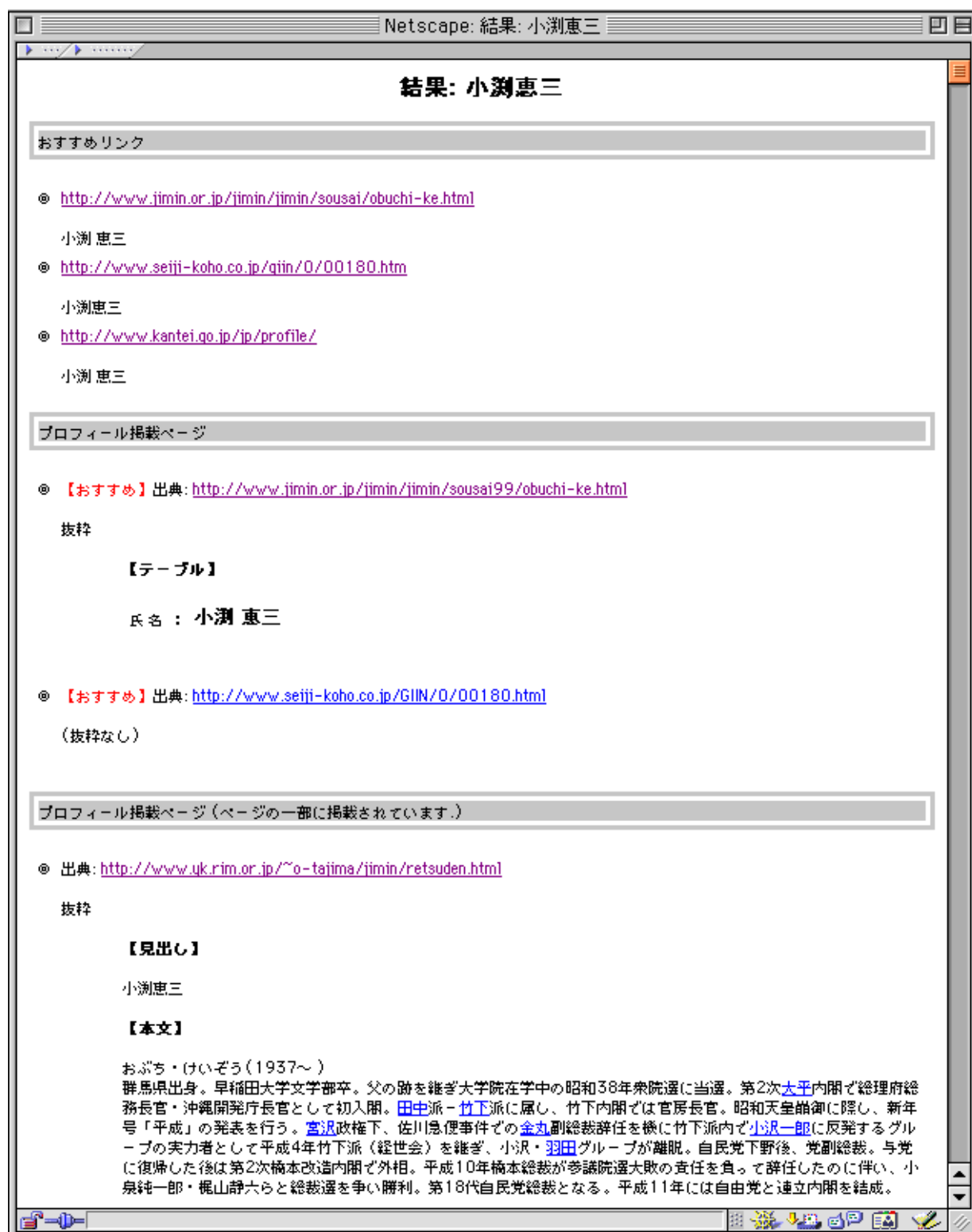


図 2.9 プロフィール収集の結果ページ

第 3 章

職名からの人物情報の収集

本章では、職名からその職業に属する人物の主要情報を収集する方法について述べる。この方法では、職業別人名録ページを情報源とし、表解析によって属性-属性値の集合として表現された人物情報を抽出する。

3.1 人物情報収集の概要

職業別人名録ページには、多数の人物の主要情報が掲載されており、その情報には、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だけでなく、その人物の職業に固有な情報も存在する。例えば、職業が政治家ならば、どの政党に属するのか、どの議会の議員であるのかという非常に重要な情報も掲載されている。本研究では、職業別人名録ページを利用することで、人物情報の収集を効率良く行う。

人物情報の収集は、次の(1) (2)の手順で行う（図 3.1 参照）。

(1) 人名録ページの収集

「政治家」などの職名を入力として、その職業の人名録ページを収集する。まず、人名録がありそうなページ（候補ページ）をワールドワイドウェブから収集し、次に、収集した候補ページの中から実際に人名録があるページのみを選別する。

(2) 表解析による人物情報の抽出

収集したページ内に存在する表形式の人名録を解析し、人物情報を抽出する。

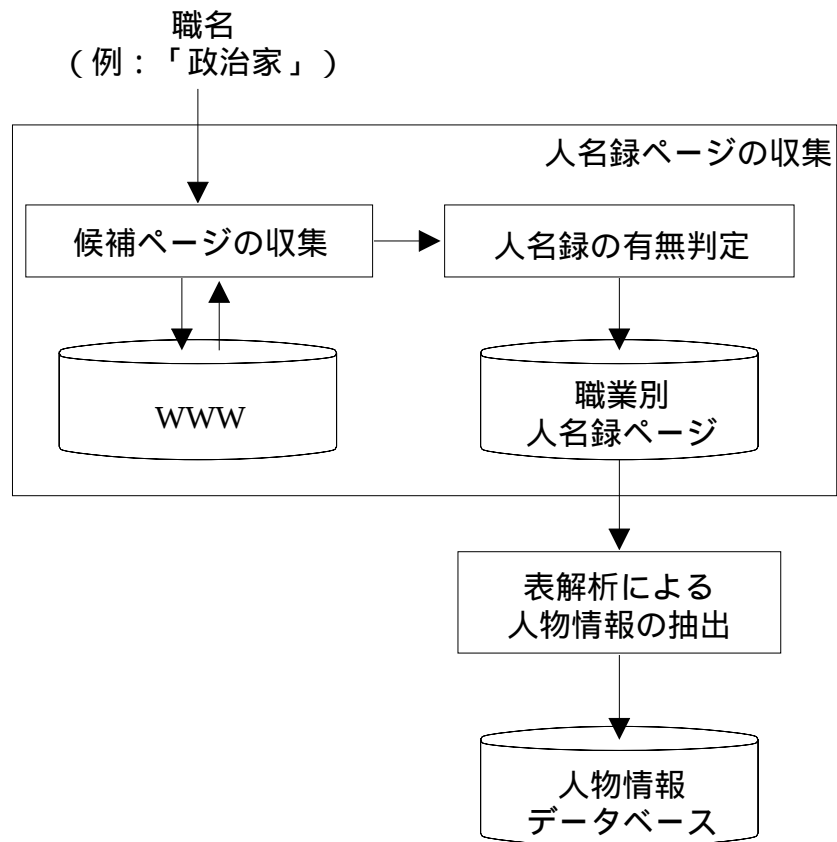


図 3.1 人物情報の収集手順

以下では ,3.2 節で人名録ページの収集方法について述べ ,3.3 節で表解析による人物情報の抽出方法について述べる .

3.2 人名録ページの収集

人名録ページの収集は ,候補ページの収集とそれらのページに対する人名録の有無判定からなる (図 3.1 参照). 候補ページの収集は ,検索エンジンとリンク情報を利用して行う .

3.2.1 検索エンジンを用いた候補ページの収集

候補ページの収集では ,職名を入力とし ,まず ,検索エンジンを用いて候補ページを

表 3.1 職名に対して定義されているクエリ

職名	クエリ
政治家	「議員名簿」, 「議員一覧」, 「議員紹介」
著述家	「一覧 and 著者 and 賞」, 「リスト and 著者 and 賞」

収集する．収集方法は，職名に対して定義されているクエリ（表 3.1 参照）を検索エンジンに入力し，検索結果として得られる URL が示すページを候補ページとして収集する．検索エンジンには，goo¹，infoseek²，LYCOS³を用い，1 クエリにつき，goo から最大 300 件，infoseek と LYCOS からはそれぞれ最大 100 件の URL を収集する．クエリは，試行錯誤によって決定した．職名が「政治家」の場合は，国会議員，および，地方議員，の人名録の収集を行うために，「議員名簿」, 「議員一覧」, 「議員紹介」の 3 つのクエリを設定した．職名が「著述家」の場合は，文学関係の受賞作品をもつ人物の人名録の収集を行うために，「一覧 and 著者 and 賞」, 「リスト and 著者 and 賞」の 2 つのクエリを設定した．

3.2.2 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は，次の 3 種類に分類できる．

(1) そのページに人名録が存在する．

(2) そのページには人名録が存在しないが，人名録へのリンクは存在する．

例を図 3.2，図 3.3 に示す．これらが示すページは，「議員名簿」を検索エンジンに入力して収集した候補ページの中に含まれていた．図 3.2 が示すページでは，「議員名簿」がハイパーリンクをもち，そのリンク先のページに議員名簿（人名録）が存在する．図 3.3 が示すページでは，「埼玉県議会」がハイパーリンクをもち，その数リンク先に議員名簿（人名録）が存在する．

(3) 人名録も人名録へのリンクもない．

¹ <http://www.goo.ne.jp/>

³ <http://japan.infoseek.com/>

² <http://www.lycos.co.jp/>



図 3.2 人名録へのリンクをもつページの例 1



図 3.3 人名録へのリンクをもつページの例 2

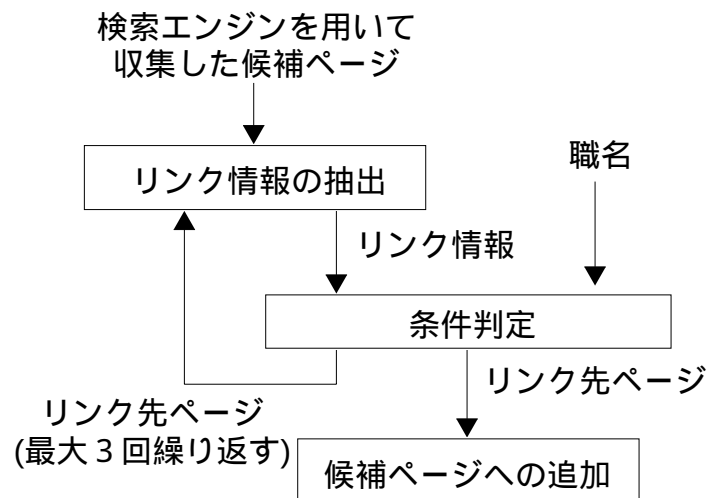


図 3.4 リンク情報を利用した候補ページの収集手順

本節では、人名録へのリンクをもったページからリンク情報を利用して、さらに、人名録ページの候補を収集する方法について述べる。ここでのリンク情報とは、アンカタグの開始タグ(<a>)から終了タグ()までの文字列と、この文字列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。以後、アンカタグの開始タグと終了タグに挟まれている文字列を、**アンカ文字列**と呼び、これにアンカタグを含めた文字列を、**アンカ**と呼ぶ。

収集手順は、次の3ステップからなる（図3.4 参照）。

(1) 候補ページから HTML タグを利用して次の方法でリンク情報を抽出する。

- (a) リストの1項目にアンカが1つだけ含まれている場合、その項目をリンク情報とする。
- (b) テーブルの、行、または、セル、にアンカが1つだけ含まれている場合、その、行、または、セル、をリンク情報とする。
- (c) HTMLソースの1行にアンカが1つだけ含まれている場合、その行をリンク情報とする。
- (d) (a)から(c)のいずれにも該当しない場合は、アンカをリンク情報とする。

表 3.2 あらかじめ定義されているキーワード

職名別に定義する キーワード (< 職名 >)	< 職種 >	職種である言葉 例：「政治家」→「議員」, 「giin」など(4語)
	< 職業関連語 >	職業と関係が深い言葉 例：「政治家」→「衆議院」, 「議会」など(6語)
職名に関係なく定 義するキーワード	< ホームページ >	ホームページを表す言葉(9語) 例：「ホームページ」, 「HomePage」 など
	< 名簿 >	名簿を表す言葉(12語) 例：「名簿」, 「リスト」, 「meibo」 など
	< 検索 >	検索サービスやリンク集を表す言葉(7語) 例：「検索」, 「リンク」, 「link」 など

英語やローマ字のキーワードは, URLと照合する.

(2) 抽出したリンク情報に対して条件判定を行う.

条件判定は, あらかじめ定義されているキーワードを用いて行う. そのキーワードを表 3.2 に示す. 抽出したリンク情報が, 以下の3つ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.

- ・アンカ文字列の末尾が< 職種 > または< 職業関連語 > である.
- ・リンク情報に< 職業関連語 > が存在し, その後に< ホームページ > が存在する.
- ・リンク情報に< 職種 > が存在し, その後に< 名簿 > または< 検索 > が存在する.

(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.

3.2.3 人名録の有無判定

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, 必ずしも人名録が存在するとは限らない. そこで, それらから人名録があるページのみを選別する.



図 3.5 50 音で分類された人名録の例

しばしば、人名録は50音で分割されることがある。例を図3.5に示す。人名録の有無判定では、このことを利用する。以下では、まず、アルゴリズムで用いる用語を定義する。

(1) 見出し

見出しとは、図や表の見出しを含め、ページ内に存在するすべての見出しを指す。見出しがハイパーリンクをもつ場合は、ページ内にその見出しの内容（テキスト、画像）がないことが多い。人名録の有無判定では、人名録の見出しが存在すれば人名録が存在すると判定するので、誤った判定をしないように、このような見出し（ハイパーリンクをもつ見出し）は除く。見出しの抽出には、HTML タグの見出しタグ（<h1> など）を利用して抽出する方法が考えられるが、HTML タグは本来の意味とは関係なく使用されることが多いので、次に示す方法を用いる。まず、HTML ソースからアンカ文字列を削除する。次に、画像タグ()以外のタグで分割した文字列（画像タグ以外のタグは含まない）のうち、34文字以下のもの（画像タグの文字数はこの数に入らない）を見出しとして抽出する。

(2) ページタイトル

ページタイトルとは、タイトルタグ(<title> , </title>)に挟まれている文字列を指す（<title> ページタイトル</title>）。

(3) 逆リンク情報

逆リンク情報とは、対象ページをリンクしているリンク情報のことを指す。例えば、図3.2に示すページには、「議員名簿」がハイパーリンクをもっており、このリンク先のページに対しては、「議員名簿」が逆リンク情報となる。

(4) 50 音文字列

50 音文字列とは、50 音分割を示す、「あいうえお順」、「50 音順」、「あ行」、「あ」、「あ～お」のような言葉のことをいう。これらの文字列は、文章中にも出現する可能性がある。文章中にも出現するものを誤って50 音文字列と判定しないように、タグで挟まれている場合、または、括弧などの記号⁴で挟まれている場合にのみ、50 音文字列と判定する。

(5) 50 音アンカ

50 音文字列がハイパーリンクをもつ場合に、50 音アンカと呼ぶ。50 音アンカは、まとめて書かれることが多い。図3.5に示す左フレームのページでも、「あ行」、「か行」などの50 音アンカがまとめて書かれている。このような特徴を利用し、50 音アンカの判定には、あらかじめ用意した50 音文字列のパターンを利用する方法以外の方法も用いる。その方法は、ページ上部からアンカ文字列を順に調べていき、あるアンカ文字列が50 音アンカと判定された場合に、その次のアンカ文字列が、ひらがな、または、カタカナ、の2文字と、「～」などの記号⁴、からなる場合は、このアンカ文字列を50 音アンカとする方法である。

(6) 50 音アンカの見出し

50 音アンカの見出しとは、何の50 音であるかを示す言葉のことを指す。例えば、図3.5における「あ行」の見出しは、「衆議院議員一覧」である。この見出しの抽出は、50 音アンカの前から順に、タグに挟まれている文字列を2つ抽出することで行う。ただし、この文字列に、50 音アンカを含めない。もし、50 音アンカが存在する場合は、さらに、タグに挟まれている文字列を抽出する。例えば、図3.5に示す左フレームのページのHTMLソース(図3.6参照)を用いると、「あ行」の見出しの抽出では、「あ行」の前で、最初にタグに挟まれている文字列は「(平成12年1月20日現在)」で

⁴ 記号は、漢字、ひらがな、カタカナ、英数字以外の文字を指す。

```

<body background=" ../images/kami_g.gif" bgcolor="#FFFFFF">
<br>
<h3 align="center">衆議院議員一覧</h3>

<p align="center"><font size="3">(平成 1 2 年 1 月 2 0 日現在)</font></p>

<p align="center"><a href="giin_a.htm" target="main">あ行</a></p>

<p align="center"><a href="giin_k.htm" target="main">か行</a></p>

<p align="center"><a href="giin_s.htm" target="main">さ行</a></p>

```

図 3.6 図 3.5 に示す左フレームのページの HTML ソース

あり，次にタグに挟まれている文字列は「衆議院議員一覧」であるから，「衆議院議員一覧 (平成 1 2 年 1 月 2 0 日現在)」が見出しとして抽出される．次の「か行」の見出しの抽出では，最初にタグに挟まれている文字列は「あ行」であるが，これは 50 音アンカであるから，次と，その次にタグに挟まれている文字列からなる「衆議院議員一覧 (平成 1 2 年 1 月 2 0 日現在)」が見出しとして抽出される．

以上で定義した用語を用いて，人名録の有無判定アルゴリズムを以下に述べる．

(1) 候補ページに人名録が存在するかどうかを次の 3 つの方法で調べる．

- (a) 見出し，ページタイトル，逆リンク情報のいずれかに，＜職名＞と＜名簿＞の両者が存在する場合，人名録が存在すると判定する．
- (b) テーブルタグの開始タグ (<table>) から終了タグ (</table>) までの文字列に，氏名を表す言葉と，職名に対して定義されている職業固有の属性名 (表 3.3 参照) が含まれている場合，表形式の人名録が存在すると判定する．氏名を表す言葉には，「氏名」，「名前」，「しめい」，「なまえ」，「名」，＜職種＞を組み合わせたものを用いる．
- (c) ページ内にハイパーリンクをもたない 50 音文字列が存在する，または，同一ページへリンクしている 50 音アンカが存在する場合，ページ内に 50 音で分割された人名録が存在すると判定する．

表 3.3 職名に対して定義されている職業固有の属性名

職名	職業固有の属性名
政治家	「政党」, 「会派」, 「党派」, 「選挙区」
著述家	「作品(名)」, 「出版社(名)」, 「出版」, 「書名」, 「タイトル」

(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録が存在すると判定する。

- ・ 50 音アンカの見出しに < 職名 > が存在する。
- ・ 候補ページ内に人名録が存在する。((1)の(a)または(b)を満たす場合)

3.3 表解析による人物情報の抽出

本節では、前節の方法で収集した人名録ページから人物情報を抽出する方法について述べる。まず、表形式の人名録から人物情報を抽出する。次に、表の見出しから職名別に定めた人物情報を抽出する。最後に、これらの情報に、職名、解析対象ページの URL とページタイトル、を加えてデータベースに格納する。

3.3.1 表からの人物情報の抽出

ワールドワイドウェブ上には、様々なレイアウトの表が存在する。これらの表は、テーブルタグ (<table>, </table>) を用いて作成した表であるかないか、属性名 (フィールド名, レコード名) を明記した表であるかないかで、大きく 4 つに分けることができる。本研究では、テーブルタグを用い、属性名が明記された表を解析対象とする。この対象となる表のうち、図 3.7 に示す表を標準形式の表とする。このような表から属性名とそれに対応する値を判定することで、属性-属性値の集合で表現された人物情報を抽出する。その際、問題となる表が、次のような表である。

場所: http://kurobe.city.kurobe.toyama.jp/life/l_gikai1.html

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	若栗3763	54-1887
金屋栄次	S19.05.15	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
中田利次	S15.01.15	無所属	三日市3342	52-0328
村椿宗輔	S08.02.22	無所属	生地吉田9650	56-9149
山内富美雄	S17.12.03	無所属	三日市3073	54-2402
山本達雄	S12.05.20	無所属	田初1199	58-1828
松原 勇	S13.01.09	無所属	生地神区349-6	56-8867
牧野和子	S18.07.20	無所属	三日市3648	52-0616
吉田重治	S14.7.15	無所属	吉田449	57-1438

図 3.7 標準形式の表の例

(1) 1つの属性名が複数のセルからなる表

例を図3.8に示す。この例の表では、「委員会 総」、「委員会 厚」などの属性名を、「委員会」と、「総」、「厚」などに分けている。

(2) 1つのセルに複数の値が存在する表

例を図3.9に示す。この例の表では、氏名と所属会派の欄に、複数の値が存在している。さらに、各々が対応している。つまり、行タグ(<tr>,</tr>)で表現された1レコードに複数レコードが存在している。

(3) ある属性の欄(行,列)に別の属性の値が存在する表

例えば、氏名の欄に、氏名だけでなく、ふりがなもあることは、よくある。他にも、図3.10に示すように、「[HP] <http://www.s-abe.or.jp/>」などは、明記されたどの属性名にも属していない。

Netscape: 秋田市 - 市議会議員名簿

場所: <http://www.city.akita.akita.jp/council/meibo.htm>

議席番号	氏名		住所	年齢	当選回数	会派	委員会					議運
							総	厚	教	建	決	
1	加賀屋千鶴子	かがやちづこ	〒011-0946 土崎港中央六丁目13番17号	37	1	共産		○				
2	明石鋭	あかしさとし	〒010-0003 東通一丁目16番22号	62	1	共産			○			
3	成沢淳子	なりさわじゅんこ	〒011-0923 飯島文京町2番3号	46	1	公明			○			
4	小林一夫	こばやしかずお	〒010-1503 下浜羽川字二十町73番地	47	1	フォーラム21	○					
5	花田清美	はなたきよみ	〒010-1421 仁井田本町一丁目15番3号	45	1	市民ク		○				
6	武藤真作	むとうしんさく	〒010-1652 豊岩豊巻字中沢33番地	46	1	市民ク			○			
7	菅原弘	すがわら	〒010-1436 大住三丁目3番	57	1	市民ク	○					

図 3.8 1つの属性名が複数のセルからなる表の例

Netscape: 議会のあらまし

Netsite: <http://www.pref.kyoto.jp/gikai/html/aramasi/aramasi2.html>

(選挙区別50音順)

選挙区(定数)		氏名	所属会派
京都市	北区(4人)	新井 進	共産
		杉谷 孝夫	公明
		武田 祥夫	府民
		田坂 幾太	自民
	上京区(3人)	田中 卓爾	府民
		林田 洋	自民
		三木 一弘	共産
	左京区(5人)	石田 昂	自民
		梅木 紀秀	共産
		北岡 千はる	府民
		澤 照美	公明
	中京区(3人)	光 永敦彦	共産
		岩田 隆夫	共産

図 3.9 1つのセルに複数の値が存在する表の例

議員名	読み仮名	選挙区	会派数値	会派TEL	会派FAX
安倍 晋三	あべ しんぞう [HP] http://www.a-site.or.jp/	山口4区	1-602	03(3508)7172	03(3508)3602
相沢 英之	あいざわ ひでゆき	鳥取2区	1-721	03(3508)7207	03(3502)3399
速沢 一郎	あいざわ いちろう [E-mail] g00055@shiofin.co.jp	岡山1区	1-405	03(3508)7105	03(3508)0319
豊知 和男	あいち かずお [HP] http://www.sashikazu.or.jp/~sashikazu/	宮城1区	2-310	03(3508)7093	03(3502)5012

図 3.10 ある属性の欄（行，列）に別の属性の値が存在する表の例

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

まず，テーブルタグに挟まれている部分に，タグ以外の文字列が1つしか含まれていない（<table><tr><td> 文字列 </td></tr></table> など）場合は，この文字列が見出しである可能性があるので，この部分のタグを削除し，見出しタグ(<h1>，</h1>)を挿入する．次に，テーブルタグが，表としてではなく，レイアウトに使用されることから，次の(a) (b)の手順で表の抽出を行う．

(a) 罫線のある表を抽出する．

もし，抽出した表の内部にテーブルタグが存在する場合は，そのテーブルに関するタグを次に示す通りに，別のものに置換，または，削除，する．行の開始タグ（<tr>）は，改行に置換する．セルの開始タグ（<td>，<th>），改行タグの開始タグ（
），段落タグの開始タグ（<p>）は，「，」に置換する．その他のタグは，削除する．

(b) 内部にテーブルタグが含まれてない表を抽出する．

抽出した表の見出しは，次の(a)から(c)の手順で抽出する．

(a) 表の内部にキャプションタグ (<caption> , </caption>) が存在する場合は , そのタグに挟まれている文字列を見出しとして抽出する .

(b) 表の 1 行目が 1 セルしかない場合は , そのセルにある値を見出しとして抽出する .

(c) 表の上部から HTML タグを利用して抽出する .

まず , 抽出した表より上の HTML ソースを , 改行タグ (
) , および , ブラウザ上で改行を行うタグ (<p> , <hr> など) で分割する . 分割された文字列のうち , 画像タグのみのものと , タグを除くと 34 文字以下になるものを対象とし , 表に近いものから , 次の手順で見出しの抽出を行う .

1. 画像タグのみの場合は , このタグから URL と , alt 属性があればその値を抽出し , これを見出しとする .

2. 見出しタグ (<h1> など) が存在すれば , このタグより後の文字列を見出しとして抽出する .

3. 文字の大きさを標準より大きくするタグ (<big> , など) が存在すれば , このタグより後の文字列を見出しとして抽出する .

5. 文字に色を付けるタグ (など) が存在すれば , このタグより後の文字列を見出しとして抽出する .

4. 太字タグ (,) が存在すれば , このタグより後の文字列を見出しとして抽出する .

(2) 表の標準化

次の(a)から(d)の手順で , 表のレイアウトを標準形式の表 (図 3.7 参照) と同様にし , 後の処理で必要となる , 表の方向 , および , フィールド名の行数 (または , レコード名の列数) の判定も行う .

(a) 次の 2 つの方法を用いて , 1 セルが複数のセルを越えていない表にする . 例えば , 図 3.10 に示す表は , 図 3.11 に示す表になる .

氏名	読み仮名	選挙区	会館部屋	会館TEL	会館FAX
あべ しんぞう		山口4区	1-602	03(3508)7172	03(3508)3602
[HP]	[HP]	[HP]	[HP]	[HP]	[HP]
http://www.s-sbe.or.jp/	http://www.s-sbe.or.jp/	http://www.s-sbe.or.jp/	http://www.s-sbe.or.jp/	http://www.s-sbe.or.jp/	http://www.s-sbe.or.jp/
あいざわ ひでゆき		鳥取2区	1-721	03(3508)7207	03(3502)3399
あいざわ いちろう		岡山1区	1-405	03(3508)7105	03(3508)0319

図 3.11 図 3.10 に示す表を標準化した表

- ・ 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 .

(b) 表の方向 (縦方向 , 横方向) を判定する .

あらかじめ用意した人物に関する属性名を用いる . 具体的には , 「氏名」 , 「年齢」など 8 語の基本的な属性名と , 職名に対して定義されている職業固有の属性名 (表 3.3 参照) の両者を用いる . まず , 1 行 2 列目から , 1 行 3 列目 , 1 行 4 列目 , と順に , 1 行目のみに対して , 人物に関する属性名と後方一致する値が存在しないかを調べていく . もし , 一致する値が存在すれば , その時点で横方向と判定する . もし , 一致する値が存在しなければ , 1 列目のみに対して同様の処理を行う . もし , この処理を行っても , 表の方向が判定できない場合は , この表の解析を放棄する .

(c) フィールド名の行数を判定する .

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 .

期別	氏名	会派	選挙区	期別	氏名	会派	選挙区
9	原 海男	自民	上野郡	3	小田 春人	自民	井原市後月郡
	門木 和郎	自民	岡山市第一		高橋 英士	公明	倉敷市都窪郡
7	榎谷 勝司	自民	岡山市第一		有木 曜	自民	赤磐郡
	小林 繁	自民	岡山市第二	2	岸本 清美	自民	鞆田郡
6	肥田 隆三郎	自民	岡山市第一		鈴木 一茂	民進	倉敷市都窪郡
	大杉 尚久	自民	高梁市		草薙 隆幸	民進	岡山市第一
5	栗山 好孝	自民	瀬口郡		長瀬 泰志	民進	倉敷市都窪郡
	小松 基勲	自民	久米郡		河本 勲	自民	湖山郡
	元原 健治	公明	岡山市第一		黒山 貴明	公明	岡山市第一
	井手 統一郎	自民	真庭郡		小野 泰弘	自民	総社市
	宮家 敦雄	自民	瀬山市		山田 祥一郎	公明	倉敷市都窪郡

図 3.12 表が結合している表の例

(d) 表が結合している場合（図 3.12 参照）は，1 つの表にする．

1 行 1 列目の値と同じ値が，横方向の表ならば 1 行目，縦方向の表ならば 1 列目，のどこに存在するかを調べる．もし，存在するものがあり，1 行 1 列目から連続して存在したものでなければ，そこを，表の結合の境とする．後は，結合している表を分割し，1 つの表にする．

(3) 標準形式の表からの人物情報の抽出

表に存在する属性名とそれに対応する値を判定し，人名別にこれらの情報をまとめることで人物情報の抽出を行う．ここでは，横方向の表の場合について述べる．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) フィールド名が氏名の位置（何行何列目）を判定する．

まず，1 行目から，1 列目，2 列目，と順に，フィールド名の部分で，氏名を表す言葉と一致する値が存在しないかを調べていく．もし，存在すれば，次に，この位置（＝位置 A）の列に存在する値を用いて，ふりがなであるかどうかを判定する．具体的には，ひらがなのみ，カタカナのみ，英字のみ，のいずれかの文字列であれば，ふりがなと判定する．ふりがなでないと判定された場合に，位置 A をフィールド名が氏名の位置とする．この方法で，位置が判定できない場合は，この表の解析を放棄する．氏名を表す言葉には，「氏名」，「名前」，

```

<フィールド名>
  <データ リンク先ページ={ “ url1 ” “ url2 ” ...}
    画像ファイル={ “ url1 ” ...}> タグのない値< /データ>
< /フィールド名>

```

図 3.13 抽出したフィールド名とその値の形式

「しめい」、「なまえ」、「名」、<職種>を組み合わせたものを用いる。

(b) 1レコードの行数を判定する。

1つの属性名が複数のセルからなる表(図3.8参照)に対応するため、次の3つの条件をすべて満たす場合のみ、1レコードの行数を1とし、他の場合は、フィールド名の行数と同じ行数とする。

- ・フィールド名の行数が2以上である
- ・フィールド名が氏名の列のフィールド名の部分には、同じフィールド名しか存在しない。
- ・フィールド名が氏名の列の最初の値から、フィールド名の行数と同じ数だけのその列の値の中に、異なる値が存在する。

(c) フィールド名と同じ順に値があるという前提で、フィールド名と値を対応させたもの(図3.13参照)を、1レコード分抽出する。ただし、以下の例外処理を行う。

1つのセルに複数の値が存在する表(図3.9参照)への対応処理

- 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。
- フィールド名が氏名であり、対応する値に改行が存在する場合は、次の1から4の手順で処理を行う。

1. 値を改行で分割する。以降、分割されたものを1つの値として扱う。
2. 値がふりがなであるかどうかを判定し、「氏名」(フィールド名)、または、「ふりがな」(フィールド名)、とこの値を対にする。(ふりがなの判定方法は、(3)の(a)で述べた方法と同様である。)
3. ふりがなが複数存在する場合は、氏名、ふりがな、の順、または、逆順、に存在することが多いことを利用し、1レコード分の氏名とふりがなを判定する。
4. 氏名が複数存在する場合は、次の方法で1レコード分を抽出する。
まず、他のフィールド名に対応する値に改行が存在するかどうかを調べる。
改行が存在しない場合は、値とフィールド名を対にし、これをすべてのレコード(氏数の数だけレコードが存在する)に含める。改行が存在する場合は、値を改行で分割し、氏名と同数の場合に限り、分割された値それぞれとフィールド名を対にし、これらを各々のレコードに含める。

ある属性の欄(行、列)に別の属性の値が存在する表(図3.10参照)への対応処理

- 抽出した1レコード内で、ある1つの値がいくつかの異なるフィールド名と対になっている場合は、それらのフィールド名は不適切と判定し、フィールド名を「？」に代える。

3.3.2 表の見出しからの人物情報の抽出

表の見出しには、その表に掲載された人物が何のカテゴリに属するのかという情報が存在する。このような情報を抽出するために、職名別に表の見出しから得られる情報を定め、抽出処理を行う。表の見出しから抽出できない場合は、ページタイトルや逆リンク情報も利用する。ここでは、実験に用いた2つの職名における抽出処理について述べる。

(1) 職名が「政治家」の場合

表の見出しから役職名(衆議院議員,石川県議員など)を抽出する。抽出方法は,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」や「予定」が含まれていないことを条件とし,次の(a)から(c)の手順で行い,抽出できた時点で処理を終了する。

(a) 表の見出し,ページタイトル,逆リンク情報の順で,次の規則に従って,役職名を判定する。

- ・「衆議院」,「衆院」,「衆議」のいずれかが存在する場合は,「衆議院議員」と判定する。
- ・「参議院」,「参院」,「参議」のいずれかが存在する場合は,「参議院議員」と判定する。
- ・「議員」,「議会」,「会議」の前に地名が存在する場合は,その地名の議員と判定する。

(b) URL に地名が存在する場合は,その地名の議員と判定する。

次に示す URL パターンを利用して,URL から地名を抽出する。

「http://www.{city, town, vill}.{市町村区名}.{都道府県名}.jp/」

または

「http://www.pref.{都道府県名}.jp/」

例えば,URL が「http://www.city.fujisawa.kanagawa.jp/gikai/meibo.html」である
とすると,「city」から市町村議員と判定し,都道府県名の「kanagawa」と市町村
区名の「fujisawa」を用いて生成した「神奈川県 fujisawa_議員」を抽出する。
なお,都道府県名は,数が少ないので,漢字表記とローマ字表記の対からなる
ファイルを用意した。これと照合してローマ字を漢字に変換する。

(c) 表の見出し,ページタイトル,逆リンク情報の順で,「議員」,「議会」,「会議」
の前に,例えば,「国会」があれば「国会議員」,「都」があれば「都道府県議員」,
「市」があれば「市町村区議員」という方法で,「国会議員」,「都道府県議員」,
「市町村区議員」のいずれであるかを判定する。

(2) 職名が「著述家」の場合

表の見出しから賞名を抽出する。抽出方法は、表の見出し、ページタイトル、逆リンク情報の順で、いずれにも「候補」や「予定」が含まれていないことを条件とし、「賞」の前にある文字列を、賞名として抽出する。

3.3.3 人物情報データベースへの格納

人物情報データベースに格納する人物情報の例を図3.14に示す。表解析で得た人物情報に、職名、解析対象ページのURLとページタイトル、を加えてデータベースに格納する。ただし、次の条件のいずれかを満たさない場合は、職名に属する人物のみからなる表でない可能性があるため、職名は人物情報に加えない。

- ・フィールド名に職名別に定義した職業固有の属性名が存在する。
- ・表の見出し、ページタイトル、逆リンク情報のいずれかに、＜職業関連語＞が存在する。

3.4 実験と検討

これまで述べてきた方法を実装したシステムを作成した。実験では、このシステムを用いる。実験を行う前に、職名別に (1)検索エンジンに入力するクエリの定義 (2)＜職種＞の定義 (3)＜職業関連語＞の定義 (4)職業固有の属性名の定義 (5)表の見出しからの情報抽出処理の用意、をする必要がある。以下では、職名別に、上述の(1)から(5)に何を設定したのかを述べ、実験の結果を述べる。

3.4.1 実験1：職名が「政治家」の場合

実験では、職名を「政治家」とし、国会議員、および、地方議員、の情報収集を行う。実験前の設定を表3.4に示す。

結果を表3.5、付録Aに示す。表3.5には、収集した人物情報を、国会議員、都道府県議員、市町村区議員、で分類して調べた結果を示す。付録Aには、これらの議員のサブカテゴリとなる役職名別に調べた結果を示す。表3.5の各欄は、以下のことを意味する。

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ります。					
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 正夫	730-0052	中区千田町三丁目6-32	082-244-0884
	公明党・ 県民会派	中田 選	730-0847	中区南入道四丁目17-9	082-295-7744
	自民	新田 篤実	730-0041	中区小町2-23	082-228-1272

データベースに格納するデータ

<林正夫>

- 表からの抽出情報 -

<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>

- 表の見出しからの抽出情報 -

<役職>広島県議員</役職>

- 職名 -

<職業>政治家</職業>

- 解析対象ページのURLとページタイトル -

<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 3.14 人物情報データベースに格納するデータの例

35

表 3.4 「政治家」における設定内容

1.検索エンジンに入力するクエリの定義	「議員名簿」, 「議員一覧」, 「議員紹介」を定義する.
2.<職種>の定義	「政治家」, 「議員」, 「seijika」, 「giin」を定義する.
3.<職業関連語>の定義	「衆議院」, 「参議院」, 「議会」, 「shugi」, 「sangi」, 「gikai」を定義する.
4.職業固有の属性名の定義	「政党」, 「会派」, 「党派」, 「選挙区」を定義する.
5.表の見出しからの情報抽出処理の用意	役職名(衆議院議員, 石川県議員など)の抽出処理を用意する.

- ・役職名：国会議員，都道府県議員，市町村区議員を代表とする．

表の見出しから抽出した役職名を用いて，人物情報を代表とした役職名のいずれかに分類する．なお，「不定」には，役職名が抽出されていなかった人物情報が該当する．

- ・URL 数：人物情報を収集した URL の数を示す．

- ・ページ数：人物情報を収集したページの数を示す．

この数は，2 つのページを比較したとき，URL が異なっている場合でも，内容が全く同じ場合は，1 と数える．

- ・役職数：システムが収集した役職の数を示す．

この役職は，役職名の欄にある役職のサブカテゴリとなるものである．

例えば，役職名が国会議員の場合は，そのサブカテゴリとして，衆議院議員，および，参議院議員，があり，システムがどちらの人物情報も収集していたら 2 となる．

- ・役職総数：実際に存在する役職の数を示す．

- ・議員定数との比較結果：システムの収集人数と議員定数を比較した結果を示す．

国会議員は，そのサブカテゴリとなる役職が 2 あり，システムはどちらの人物情報も収集した．ただし，どちらも定員以上の人数が見つかった（表 3.6 参照）．この原因を調べたところ，字体の違い，および，旧人名録と新人名録の混合，がその原因となっていた．

表 3.5 役職名別に調べた結果

役職名	URL数	ページ数	役職数	役職総数	議員定数との比較結果
1.国会議員	50	37	2	2	定数以上が2件
2.都道府県議員	30	26	13	47 (人手18)	定数以下が 1 件
3.市町村議員	24	23	22	3380	一致
不定	7	7	-	-	-
合計	111	86	37	3429	-

表 3.6 議員定数との比較

役職名	収集人数	定数
衆議院議員	546	500
参議院議員	254	252

都道府県議員は、そのサブカテゴリとなる役職が 47 あり、本システムで利用した検索エンジン (goo, infoseek, LYCOS) を用いて著者が探したところ、そのうち 18 に対して議員名簿があることが確認された。システムはそのうち 13 に対して人物情報を収集した。この 13 の役職別に、収集人数と、議員定数を比較したところ、定数以下のものが 1 存在した。この原因は、その役職の人名録をシステムがすべて収集していなかったことにある。また、ウェブ上での存在が確認されている残り 5 つの役職に属する人物情報が全く収集されていなかった原因を調べた。その結果を次の(1) (2)に示す。

(1) 人名録ページが収集されていなかった (役職数 : 3)

人名録へのリンクをもつページは収集していたが、そのリンク先にある人名録ページを収集していなかった。これは、人名録が選挙区で分割されているため、本システムは収集できなかった。本システムは、50 音で分割されている場合のみを対象としている。

(2) 人名録ページから情報抽出されていなかった (役職数 : 2)

人名録がテーブルタグを用いないで作成した表であったため抽出できなかった。

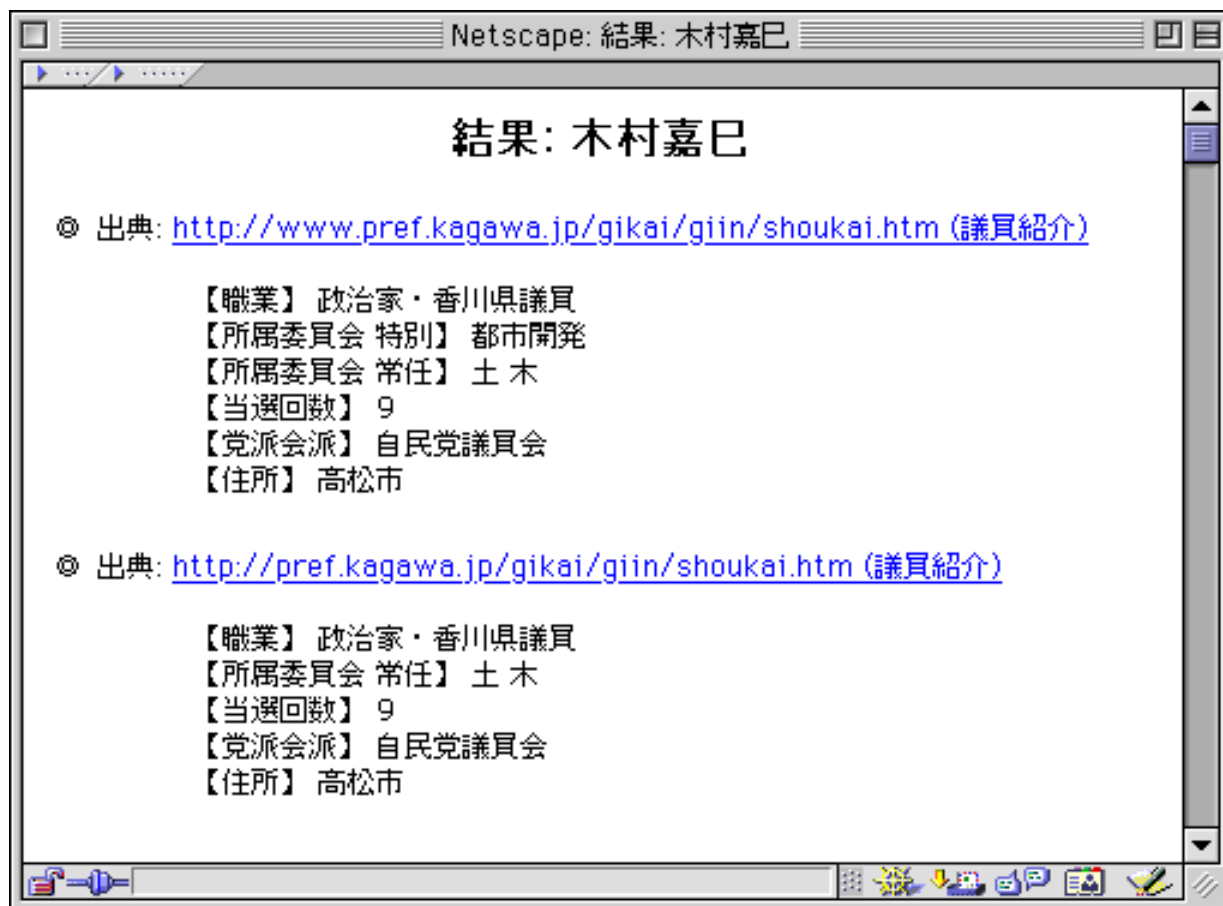


図 3.15 抽出した人物情報の中で人名が「木村嘉巳」であるもの

市町村区議員は、そのサブカテゴリとなる役職名が3380あり、そのうち21に対して人物情報を収集した。

表解析は、都道府県議員の人名録を参考に作成したものであるが、他の議員においても表解析から得た人物情報に属性 - 属性値の対応が不適切なものは見られなかった。

抽出した人物情報には、人物の名前、職業だけでなく、政治家である場合に重要な情報である政党なども存在した。例を図 3.15 に示す。

表 3.7 「著述家」における設定内容

1.検索エンジンに入力するクエリの定義	「一覧 and 著者 and 賞」，「リスト and 著者 and 賞」を定義する．
2.<職種>の定義	「著述家」，「作家」，「著者」，「受賞者」を定義する．
3.<職業関連語>の定義	「文学」，「小説」，「エッセイ」，「ミステリー」，「賞」を定義する．
4.職業固有の属性名の定義	「作品(名)」，「出版社(名)」，「出版」，「書名」，「タイトル」を定義する．
5.表の見出しからの情報抽出処理の用意	賞名（芥川賞，直木賞など）の抽出処理を用意する．

3.4.2 実験 2：職名が「著述家」の場合

実験では，文学関係の受賞作品をもつ人物の情報収集を行う．実験前の設定を表 3.7 に示す．

結果を表 3.8 に示す．システムが収集した人物情報を調べたところ，23 の人名録ページから 942 人の人物情報を収集していた．942 人の人物情報を調べたところ，次に示す問題が見つかった．

- ・著者フィールドにある値を氏名とするため，編集局名などが氏名と判定される．
- ・表解析は，属性名と同じ順に値があるという前提で行うため，この前提に反する表を解析した場合，属性 - 属性値の対応が不適切なものが抽出される．例えば，図 3.16 に示す表を解析すると，「選考委員」と「読者賞 桜子は帰って来たか 麗羅」を対応させたものが抽出される．このことが原因で，表からすべての人物情報が抽出されていなかったものが，実験では 4 存在した（表 3.8 参照）．

また，1 つの賞に対する人名録ページの数，1 ページである場合がほとんどであった．これは，賞名を与えないで人名録の収集を行ったためで，賞名を与えることでその賞に関する人名録ページをより多く収集できるだろう．

抽出した人物情報には，人物の名前，職業だけでなく，著述家である場合に重要な情報である受賞作品とその賞名なども存在した．例を図 3.17 に示す．

表 3.8 賞名別に調べた結果

賞名	収集人数	ページ数	表からの抽出状況
1.読売文学賞	38	1	すべて
2.直木三十五賞	101	1	すべて
3.谷崎潤一郎賞	30	1	すべて
4.鮎川哲也賞	16	1	一部
5.サントリーミステリー大賞	26	1	一部
6.小説大賞	261	4	すべて
7.吉川英治文学新人賞	30	1	すべて
8.日本推理作家協会賞	63	1	すべて
9.芥川龍之介賞	89	1	すべて
10.日本推理サスペンス大賞	18	1	一部
11.吉川英治文学賞	27	1	すべて
12.横溝正史賞	38	1	一部
13.山本七平賞	12	1	すべて
14.野間文芸賞	40	1	すべて
15.江戸川乱歩賞	44	1	すべて
(賞名なし)	109	5	すべて
合計	942	23	-

回	年度	受賞名	受賞作品	著者	出版社	出版年
第1回	昭和58年 (1983年)	選考委員				
		大賞	虹へ、アヴァンチュール	鷹羽十九哉	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	黒川博行	文藝春秋社	84/09
		阿川弘之、関高健、小松左京、田辺聖子、都筑道夫				

(<http://www.inv.co.jp/baba/book/list/suntry.html>)

図 3.16 フィールド名と同じ順に値がない表の例

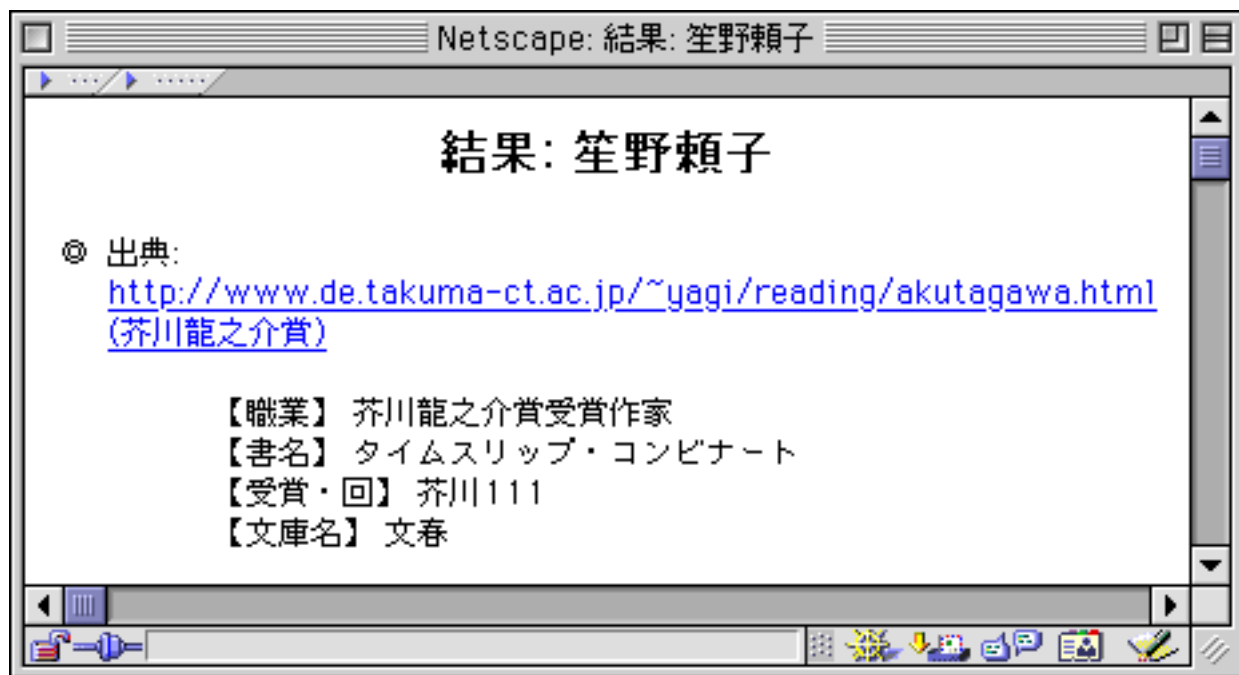


図 3.17 抽出した人物情報の中で人名が「笙野頼子」であるもの

3.4.3 検討

人物の名前と、職業だけでなく、その職業に固有な情報も収集できており、職業別人名録を利用した方法は人物情報を収集する方法として有効であると言える。

職業別人名録は、表形式で書かれる場合、フィールド名が明記されていることが多い。したがって、他の職業においても人物情報の収集が可能である。ただし、表形式で書かれるのは、人数が少ない場合である。人数が多い場合は、データベースを用いることが多い。また、より多くの人物情報を掲載するために、それぞれの人物に対してページを用意する人名録も少なくない。より多くの人物情報を収集するためには、表解析以外の方法も必要である。

職業別人名録が50音で分割されることは、どの職業でも有り得ることであるが、職業によってよく使用される分割の仕方も存在する。例えば、研究者総覧(人名録)ならば学部・学科名で分割されることが非常に多い。より多くの人名録を収集するためには、50音以外で分割された人名録を収集する必要がある。

第 4 章

人名からの人物情報の収集

本章では、人名からその人物の主要情報を収集する方法について述べる。この方法では、人物の主要情報がテキストで記述されているページを情報源とし、レイアウト解析を行って、テキスト形式の人物情報を抽出する。

4.1 人物情報収集の概要

情報源とするテキスト形式の人物情報(プロフィール)が存在するページは、1 名のプロフィールをページ全体に記述したページ(人物紹介ページ)と、1 名または複数名のプロフィールをページの一部に記述したページに分けられる。ある人物のプロフィールを収集する場合、前者のページが見つかった場合は、このページの URL を収集する。後者のページが見つかった場合は、求める人物のプロフィールのみを抽出する。これを実現するためには、人物紹介ページの判定と、プロフィールの抽出、が必要となる。

本研究では、人物紹介ページの判定、および、プロフィールの抽出、をそこでよく用いられているレイアウトに着目して行う。例えば、人物紹介ページでは、ページの始めに人名の見出しが存在したり、この見出しを罫線で挟んだりするレイアウトがよく見られる。プロフィールでは、人名の見出しの後に空行があり、その次の行から人物の紹介を記述する、あるいは、人名の見出しの後の行をインデントして人物の紹介を記述する、などのレイアウトがよく見られる。複数名のプロフィールが列挙されている場合は、これらのプロフィールの間に、空行、または、罫線、があることが多い。

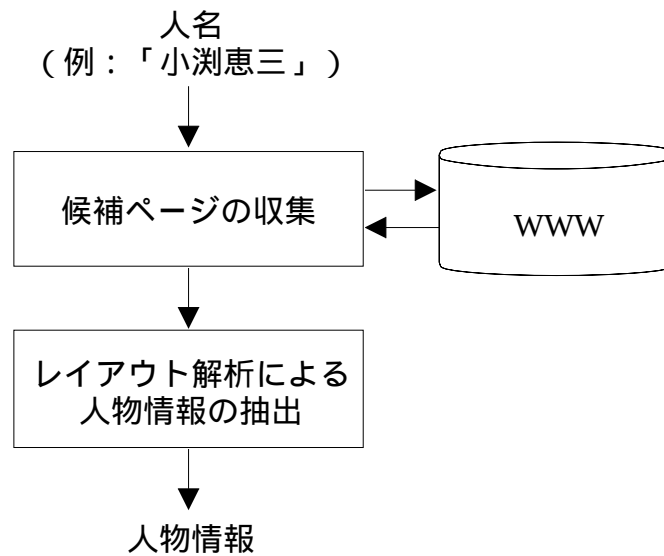


図 4.1 人物情報の収集手順

人物情報の収集は、次の(1) (2)の手順で行う（図 4.1 参照）。

(1) 候補ページの収集

人名を入力とし、その人物のプロフィールがありそうなページ(候補ページ)をワールドワイドウェブから収集する。

(2) レイアウト解析による人物情報の抽出

収集したページからレイアウト情報を利用してプロフィールを抽出する。さらに、人物紹介ページであるかの判定も行う。

以下では、4.2節で候補ページの収集方法について述べ、4.3節でレイアウト解析による人物情報の抽出方法について述べる。

4.2 候補ページの収集

候補ページの収集では、次の2つの方法を用いる。

(1) 事前に収集した職業別人名録ページからの収集

3章で述べた方法で収集した職業別人名録ページから人名が含まれているページを候補ページとして収集する。

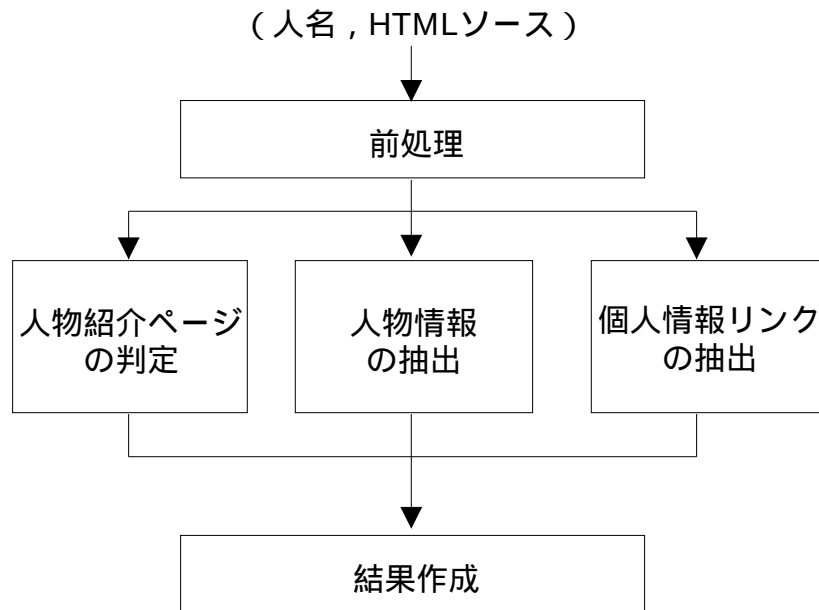


図 4.2 人物情報の抽出手順

(2) 検索エンジンを用いた収集

まず、人名から検索エンジンへ入力するクエリを作成する。次に、このクエリを検索エンジンに入力して検索結果であるURLが示すページを候補ページとして収集する。検索エンジンは、goo、infoseek、LYCOSを用い、最大検索件数は各々10件とする。検索エンジンに入力するクエリは、「人名 and 生まれ」、「人名 and 生れ」、「人名 and 出身」、「人名 and 略歴」、「人名」の5つである。

4.3 レイアウト解析による人物情報の抽出

本節では、前節の方法で収集した候補ページに対して、レイアウト情報を示すタグを挿入し、これを利用することでプロフィールを抽出する方法について述べる。1ページからプロフィールを抽出する手順を図4.2に示す。まず、HTMLソースにレイアウト情報を示すタグを挿入するなどの前処理を行う。次に、レイアウト情報が挿入されたHTMLソースと人名を入力として、次の3つの処理を行う。

- (1) 人物紹介ページであるかどうかを判定する．
- (2) 人名の見出しをもつプロフィールを抽出する．
- (3) 個人情報が記載されたページへのリンク情報（個人情報リンク）を抽出する．

最後に，これらの結果を URL と共に保存する．以上の手順で，すべての候補ページに対して同様の処理を行う．以下に，各処理の方法について述べる．

4.3.1 前処理

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利用できる．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．そこで，HTML ソースに対して次に示す(1) (2)の処理を行う．

(1) 次の方法で，表以外の用途で使用されているテーブルタグを削除する．

- (a) タグ以外の文字列が 1 つしか含まれていないテーブルは，この文字列を見出しと判定し，見出しタグを挿入する．テーブル関係のタグ（<table>，<tr>，<td> など）は，削除する．
- (b) アンカのみからなるテーブルは，ナビゲーションと判定し，各アンカの前にリストタグ（）を挿入する．テーブルタグは，ul タグに置換し，その他のテーブル関係のタグは，削除する．
- (c) アンカのみからなる列と，それらのリンク先のページの説明文のみからなる列をもつテーブルは，リンク集と判定し，各アンカの前にリストタグを挿入する．テーブルタグは，ul タグに置換し，その他のテーブル関係のタグは，削除する．
- (d) 次のいずれかに該当する場合は，ページのレイアウトにテーブルタグを使用していると判定し，セル内の文字列の始めと終わりに段落タグ（<p>，</p>）を挿入する．テーブル関係のタグは，削除する．

- ・ 画像のみからなるテーブル

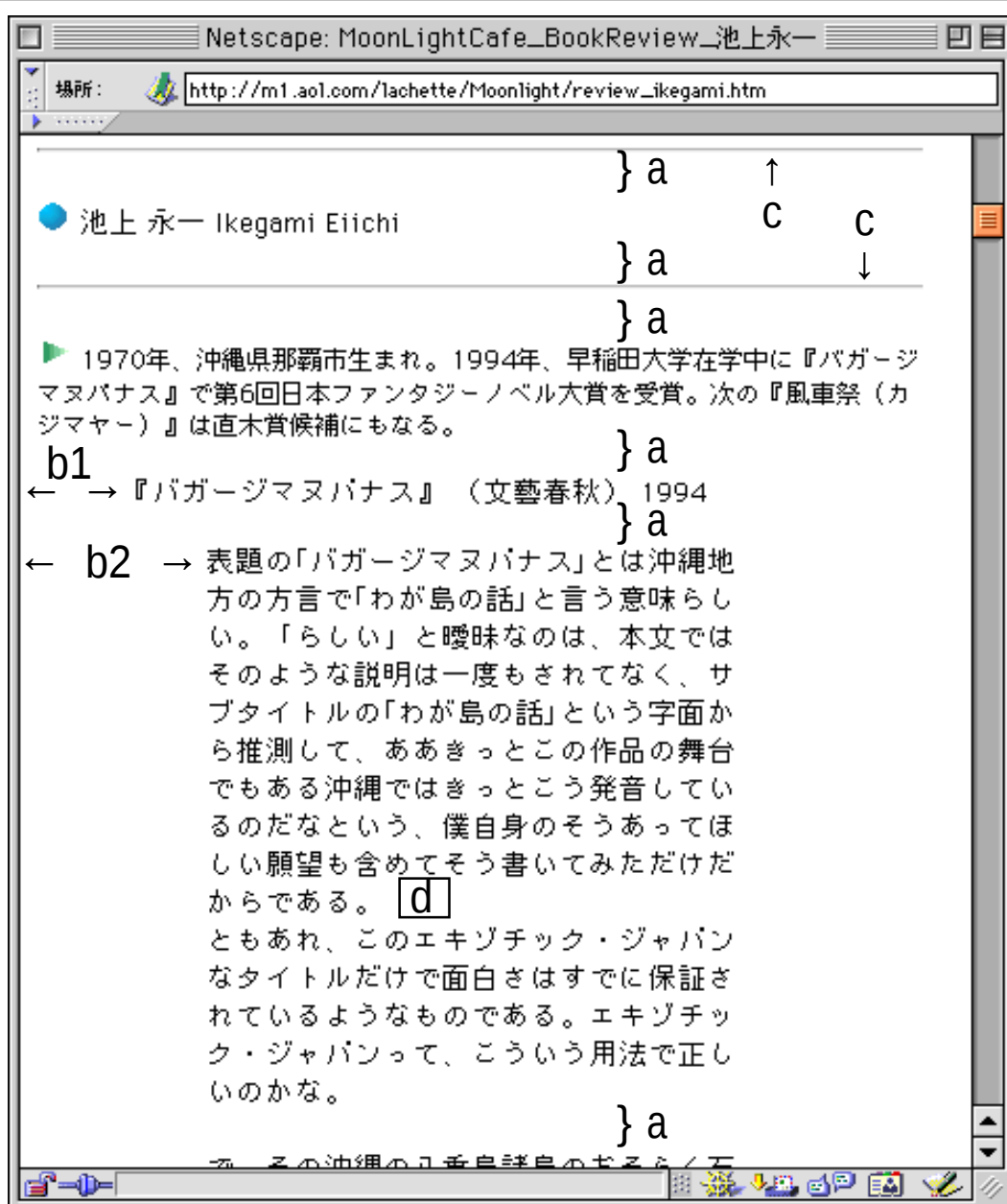
- ・画像のみからなる行(または列)と画像を参照している文のみからなる行(または列)をもつテーブル
- ・リストタグや見出しタグが含まれているテーブル
- ・「です。」などの文末表現の言葉が含まれているテーブル
- ・(a)から(c)と上述のいずれかに該当するテーブルが含まれているテーブル

(2) HTMLソースにレイアウト情報を示す4種類のタグを挿入する。4種類のタグとは、空行タグ、改行タグ、罫線タグ、深さ付きのインデントタグである。これら4種類のタグが示すレイアウト情報を図4.3に示す。空行タグは、ブラウザ表示の際に、そこに空行が存在していることを示す。改行タグは、ブラウザ表示の際に、そこが強制改行されていることを示す。罫線タグは、ブラウザ表示の際に、そこに罫線が存在していることを示す。深さ付きのインデントタグは、ブラウザ表示の際に、行の始めがインデントされていることを示し、その深さ別に開始タグと終了タグを用いる。ある深さのインデントが始まる最初の行の前に開始タグを挿入し、その深さが終了する最後の行の末尾に終了タグを挿入する。深さは、数字で表し、数が大きくなるほど深いことを示す。

4.3.2 人物紹介ページの判定

レイアウト情報が挿入されたHTMLソースと人名を入力とし、人物紹介ページであるかを判定する。判定条件には、あらかじめ定義されているキーワードを用いる。そのキーワードを表4.1に示す。次の5つの条件のいずれかを満たす場合、人物紹介ページであると判定する。

- (1) ページタイトルに人名と<プロフィール>が存在する。
- (2) ページタイトルに人名が存在し、ページの本文に<略歴>が存在する。
- (3) ブラウザ表示の1行目に人名が存在し、それ以降のテキストに<略歴>が存在する。
- (4) 最初の罫線の後の行に人名が存在し、それ以降のテキストに<略歴>が存在する。
- (5) 属性-属性値集合の人物情報がページ内に1人分のみ存在する。



- a... 空行
- b1... インデント（深さ1）
- b2... インデント（深さ2）
- c... 罫線
- d... 改行

図 4.3 レイアウト情報

表 4.1 定義されているキーワード

キーワードの種類	定義されているキーワード
<プロフィール>	プロフィールを表す言葉（「自己紹介」など6語）
<氏名>	氏名を表す言葉（「氏名」，「名前」など4語）
<略歴>	略歴を表す言葉（「略歴」，「経歴」など6語）
<属性名>	<氏名>と<略歴>に定義されているキーワードと， 「生年月日」，「年令」，「年齢」

まず，「<氏名>，セルタグ(<TD>)，人名」，または，「<氏名>，区切り文字¹，人名」，の並びからなる文字列がHTMLソース内に存在するかを調べる．存在した場合は，この文字列に対して，次の2つの条件を共に満たす場合に，ページ内に人物情報が1人分のみ存在すると判定する．

- ・この文字列の人名の部分が異なる文字列がHTMLソース内に存在しない．
- ・この文字列の<氏名>の部分を他の<属性名>にした文字列がHTMLソース内に存在する

4.3.3 テキストからの人物情報の抽出

レイアウト情報が挿入されたHTMLソースと人名を入力とし，人名の見出しをもつプロフィールを抽出する．抽出方法は，次の(1) (2)に示す．

- (1) 「人名の見出し，レイアウト情報を示すタグ，人物紹介のテキスト」の並びからなる人物情報の抽出

抽出対象となる人物情報の例を図4.4に示す．抽出対象となる人物情報は，人名の見出しの後に，空行，改行，罫線，インデント，のいずれかが存在し，その後のテキストに「生まれ」などの人物紹介によく使用されるキーワード（人物紹介キーワー

¹区切り文字は，漢字，ひらがな，カタカナ，英数字以外の文字を指す．

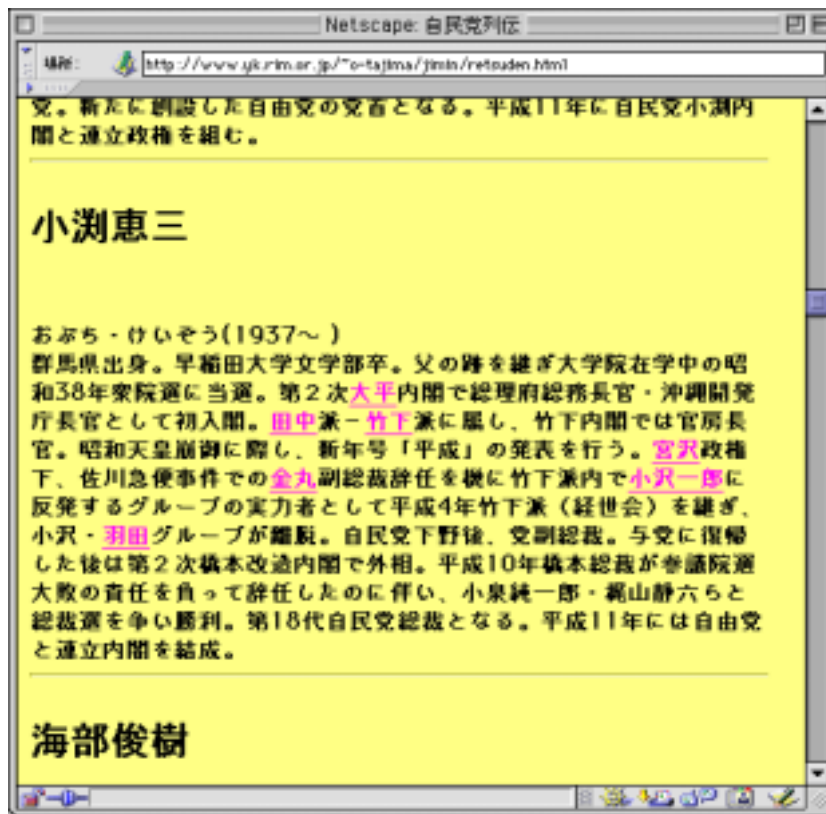


図 4.4 「人名の見出し，空行，人物紹介のテキスト」
の並びからなる人物情報の例

ド)が存在するものである．この人物情報を抽出するためには，ページ内の見出しと，その本文の領域，を判定する必要がある．以下に，抽出手順を示す．

(a) HTML ソースに確信度付きの見出しタグを，次に示す 3 つの情報を利用して，挿入する．

- ・レイアウト情報：インデントの位置と深さ．罫線，空行，改行の位置．
- ・書式情報：文字の色，太さ，大きさ．
- ・テキスト情報：文字列の長さ．「・」などの記号との区別．

まず，空行で分割したもののうち 60 文字を超えないものと，レイアウト情報を示すタグで分割したもののうち 60 文字を超えないもの，を見出しの候補とす

表 4.2 確信度の付け方

書式条件 レイアウト条件	見出しが強調 されている .	見出しが強調 されていない .
見出しの前後に , 空行 , 罫線 , インデント のいずれかがある .	5	4
見出しの後に , 空行 , 罫線 , インデント のいずれもない .	3	

見出しの前に , 罫線 , インデントのいずれかがある場合は 2 ,
ない場合は , 1 とする .

る . 次に , 候補の見出しに対して , レイアウト情報と , 書式情報の 2 点から , 表 4.2 に従い , 確信度を付ける . 表 4.2 の見方について述べる . まず , 書式条件の欄の「見出しが強調されている .」というのは , 見出しが , 黒以外の色である , 標準の大きさより大きい , 太字である , のいずれかに該当する場合を指す . これらのいずれかに該当する場合は , 「見出しが強調されている」の列を見る . いずれにも該当しない場合は , 「見出しが強調されていない」の列を見る . 次に , レイアウト条件の欄を見る . 2 つの条件が書かれているが , いずれかに該当する場合は , その行を見て , 先程の書式条件で決まった列と交わるところが候補の見出しに付ける確信度となる . いずれにも該当しない場合は , 確信度 0 とする . 確信度は , 0 から 5 までの数字で表し , 見出しである可能性が高いものほど数が大きくなる . 最後に , 確信度が 0 以上の候補の見出しは , 見出しであると判定する .

(b) 見出しに人名が存在する場合は , その本文を抽出する .

本文の領域は , 人名をもつ見出しの次の行から , 人名をもつ見出しがあるところと同じインデントの深さを基本とし , 人名をもつ見出しの確信度と , 同じ , または , 高い , 見出しがあれば , その前の行までとする . なお , 人名をもつ見出しの次の行からインデントの深さが深くなる場合は , その深い部分を本文の領域とする .

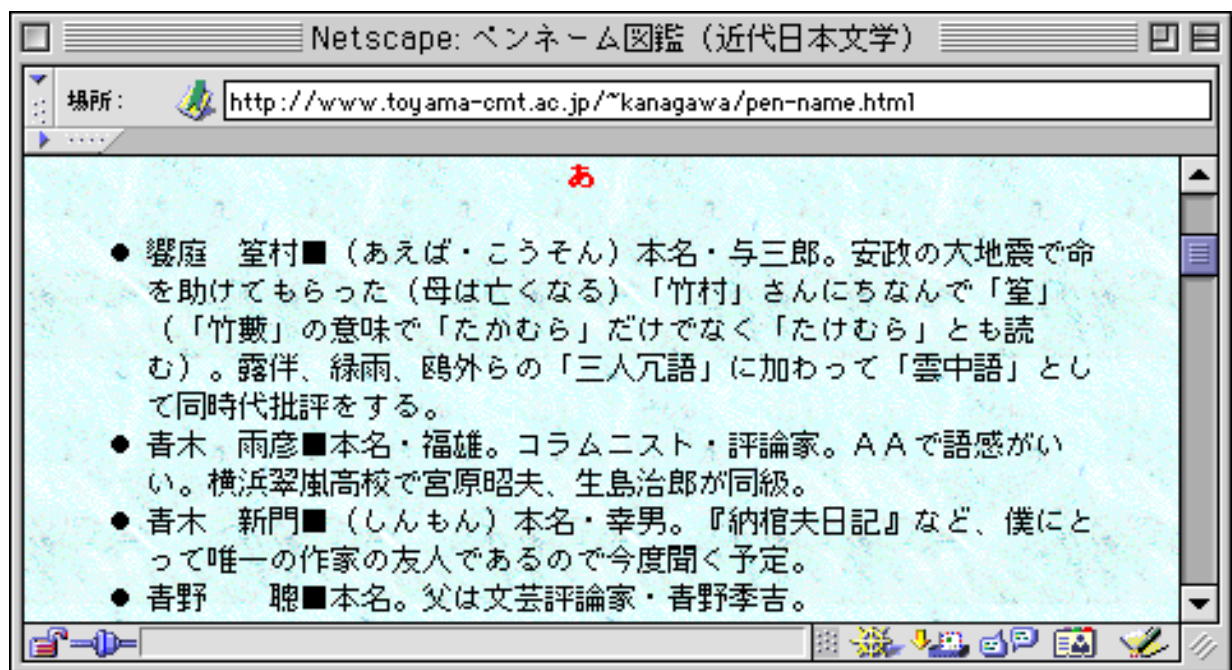


図 4.5 「人名の見出し，区切り文字，人物紹介のテキスト」
の並びからなる人物情報の例

(c) 抽出した本文に，人物紹介キーワードが存在するかどうかを調べる．

人物紹介キーワードとして，「生まれ」，「生れ」，「出身」，「本名」を用いる．これらのいずれかが本文に存在する場合，見出しと共に人物情報として抽出する．

(2) 「人名の見出し，区切り文字，人物紹介のテキスト」の並びからなる人物情報の抽出
抽出対象となる人物情報の例を図4.5に示す．抽出対象となる人物情報は，リストの1項目としてレイアウトされており，1項目の始めに人名の見出しが存在し，その後のテキストに人物紹介キーワードが存在するものである．抽出は，次の(1) (2)の手順で行う．

(1) リストの1項目を抽出する．

リストの1項目は，HTMLのリストタグを利用して抽出したもの以外に，行頭が「・」などの記号から始まるものも，次の改行タグまでとして，用いる．

(2) 抽出したリストの項目のうち ,次に示す 2 つの条件を共に満たすものを ,求める人物情報として抽出する .

- ・ 1 項目の始めが ,人名 ,名詞や接尾辞である文字列 ,区切り文字 ,を組み合わせた文字列である (名詞や接尾辞の判定には ,日本語形態素解析システム JUMAN[12]を使用した .)
- ・ 人物紹介キーワードが存在する (人物紹介キーワードは (1)の(c)に示したものと同様のものである .)

4.3.4 個人情報リンクの抽出

レイアウト情報が挿入されたHTMLソースと人名を入力とし ,HTMLソースに存在するリンク情報が ,次に示す 3 つの条件のいずれかを満たす場合 ,個人情報リンクとして抽出する .

- ・ アンカの末尾が人名である
- ・ リンク情報に人名と<プロフィール>が存在する
- ・ リンク情報に人名と<ホームページ> (表 3.2 参照) が存在する

4.3.5 結果作成

4.3.1項から 4.3.4項で述べた方法で得た情報と ,対象ページのURL からなる結果を作成する .この結果を利用することで ,利用者に人物紹介ページのリンク集 ,プロフィール集 ,個人情報リンク集を提供することが可能となる .

4.4 実験と検討

これまで述べてきた方法を実装したシステムを作成した .このシステムは ,人名を入力とするが ,その人物の職業も指定した場合は ,事前に収集した職業別人名録ページから候補ページの収集を行い ,人名のみの場合は ,検索エンジンを用いて候補ページの収

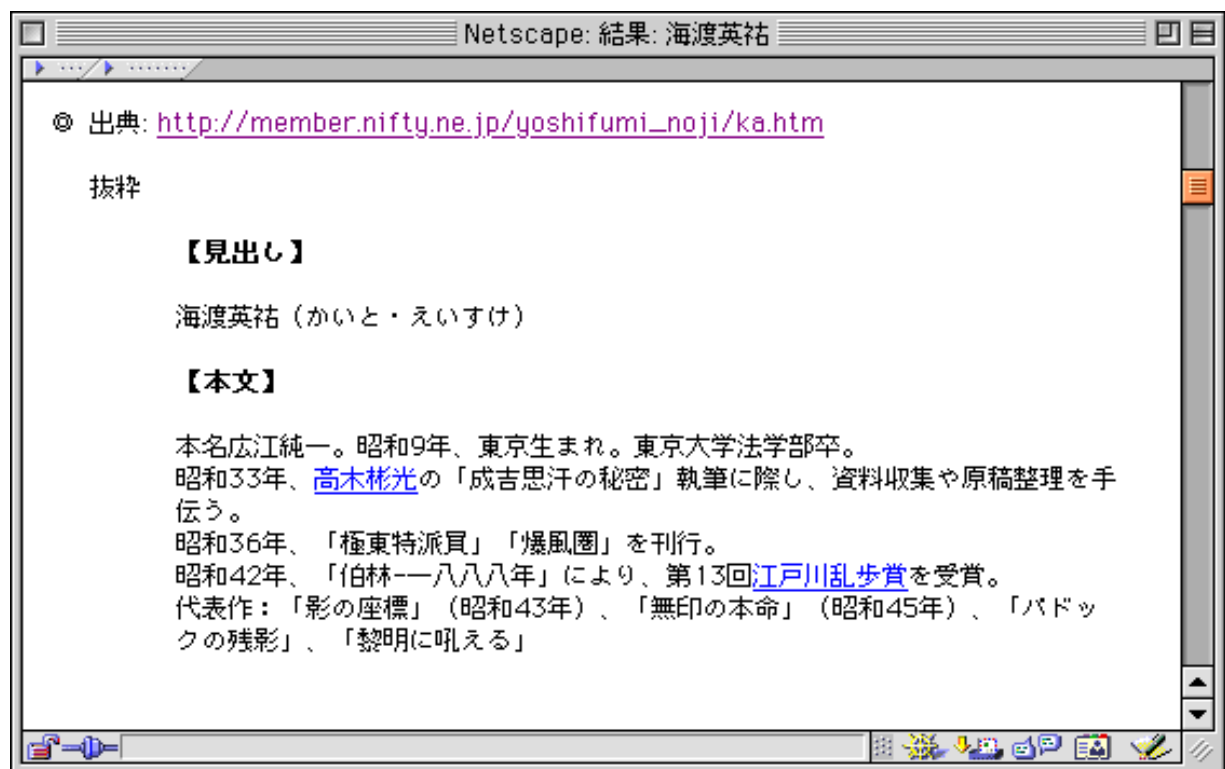


図 4.6 不要な部分を含まない正解プロフィールの例

集を行う。実験では、このシステムを用い、人名の見出しをもつプロフィールの収集状況について調べる。その際、抽出したプロフィールの正解、不正解の判断を行うが、これは、著者が行ったものである。

正解プロフィールの例を図 4.6 に示す。正解としたプロフィールは、対象人物のプロフィールが含まれているものである。つまり、不要な他のテキストが含まれていても正解とした。ただし、対象人物のプロフィールは、数文以上からなるものであり、対象人物の紹介を目的として記述されたテキストとする。例えば、図 4.7 に示すプロフィールは、不正解とする。不要な他のテキストが含まれている場合には、対象人物以外のプロフィールも含まれている場合と、プロフィール以外のテキストも含まれている場合、がある。各々の例を図 4.8、図 4.9 に示す。

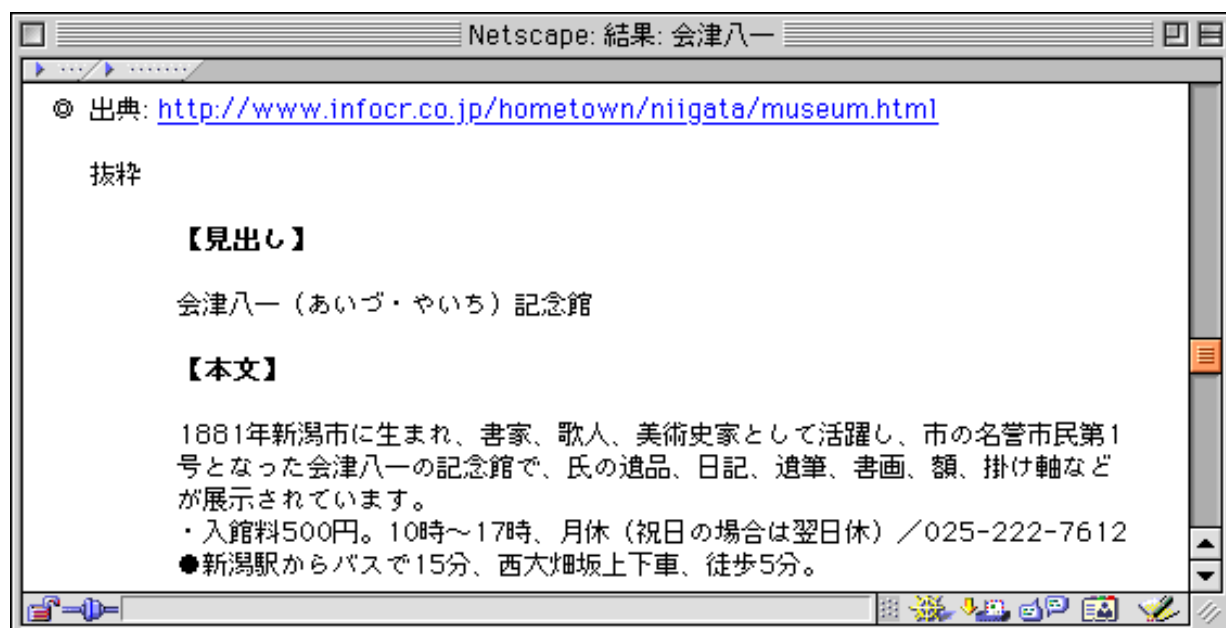


図 4.7 不正解プロフィールの例

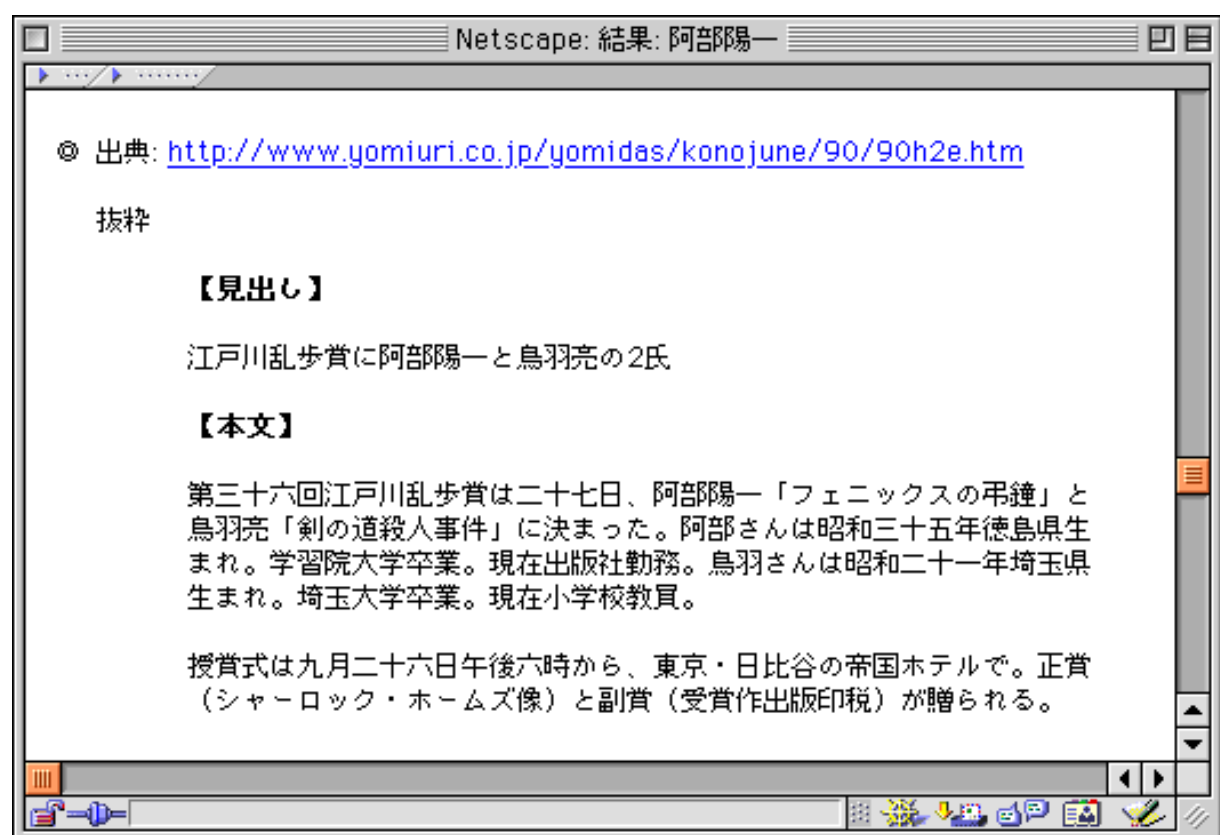


図 4.8 対象人物以外のプロフィールも存在する正解プロフィールの例

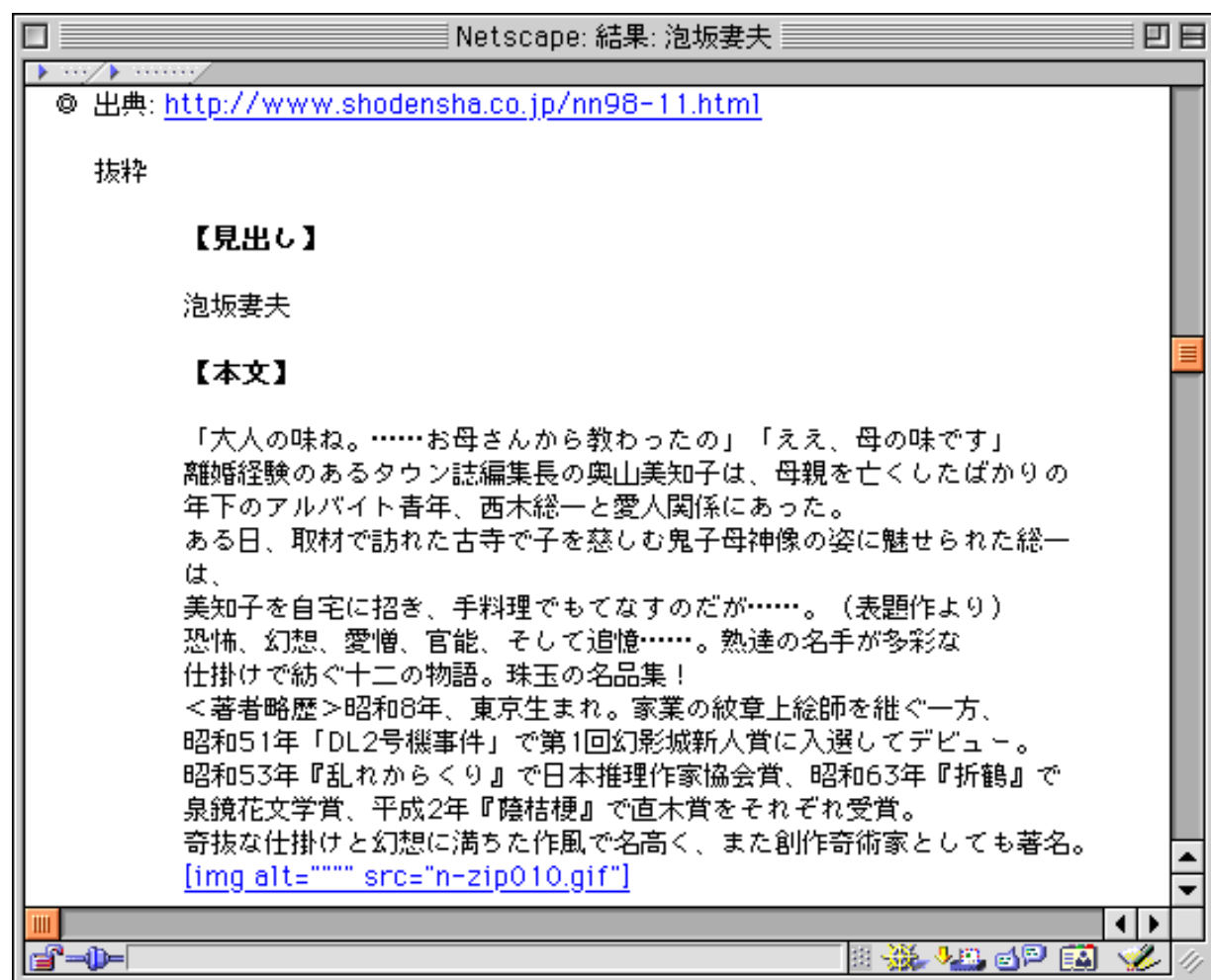


図 4.9 プロフィール以外のテキストも含む正解プロフィールの例

表 4.3 実験 1 の人物情報の収集状況（参議院議員 50 名）

収集状況	正解プロフィールを含む	不正解プロフィールのみ	合計
プロフィールあり	3名	1名	4名
他の情報のみ	-	-	10名
情報なし	-	-	36名

個人情報リンクのみである人物が9名存在した。残り1名は、個人情報リンクと人物紹介ページのURLを収集していた。

表 4.4 実験 1 の収集プロフィールの状況（参議院議員 50 名）

不要な部分	正解プロフィール	不正解プロフィール
ない	2件	-
ある	0件	-
合計	2件	1件

4.4.1 実験 1：人名と職業を指定した場合

実験では、人名と職名を入力とし、その職業の人名録ページから人物情報を収集する。人名録ページは、3章の実験で収集したものをうい、人名は、3章の実験で収集したもののうち、ランダムに選んだ、参議院議員 50 名と、著述家 128 名、を用いた。

参議院議員 50 名の結果を表 4.3、表 4.4 に示す。50 名のうち 4 名からプロフィールを収集した。ただし、4 名のうち 1 名は不正解のプロフィールしか存在しなかった。収集したプロフィールは、全部で 3 件あり、そのうち正解は 2 件であった。収集件数がかなり少ないのは、情報源とした人名録ページに存在する人物情報は、表形式のものがほとんどであるためである。

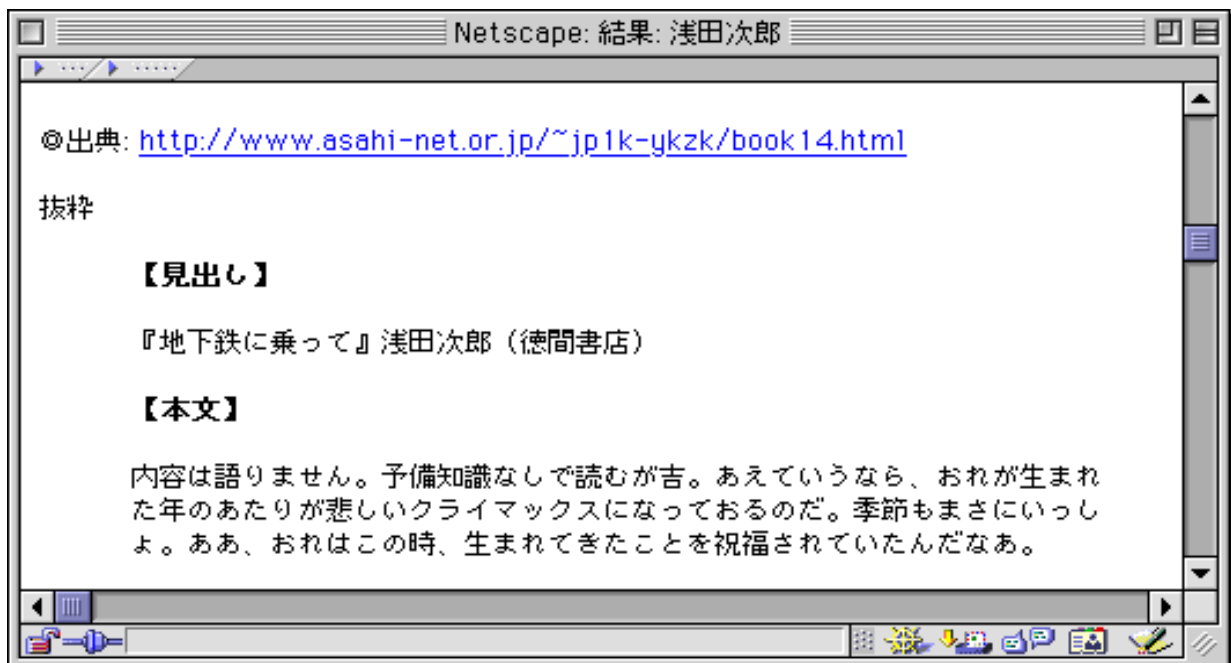
著述家 128 名の結果を表 4.5、表 4.6 に示す。128 名のうち 34 名からプロフィールを収集した。ただし、34 名のうち 3 名は不正解のプロフィールしか存在しなかった。したがって、正解プロフィールを収集していた人数は、全体の 24% であった。収集したプロフィールは、全部で 69 件あり、そのうち正解は 59 件であった。

表 4.5 実験 1 の人物情報の収集状況（著述家 128 名）

収集状況	正解プロフィールを含む	不正解プロフィールのみ	合計
プロフィールあり	31名	3名	34名
個人情報リンクのみ	-	-	17名
情報なし	-	-	77名

表 4.6 実験 1 の収集プロフィールの状況（著述家 128 名）

不要な部分	正解プロフィール	不正解プロフィール
ない	53件	-
ある	6件	-
合計	59件	10件



（本文に含まれている「生まれ」は「浅田次郎」のプロフィールとは関係がない）

図 4.10 不正解プロフィールの例

プロフィールとして不適切なものを抽出した原因は、人名をもつ見出しの本文にその人物を紹介したテキストが存在するかどうかの判定方法にある。本システムは、「生まれ」などの人物紹介によく使用されるキーワードが含まれているかいないかで判定している。そのため、含まれていたキーワードが人物紹介とは関係がないものであっても、人物を紹介したテキストが存在すると判定してしまう。例えば、図4.10に示すものがプロフィールとして抽出されていた。

4.4.2 実験 2：人名のみを指定した場合

実験では、人名を入力とし、検索エンジンを用いて収集した候補ページから人物情報を収集する。人名は、前項と同様のものを用いる。なお、ここで収集する人物情報は、職名に関係なく収集した人物情報である。つまり、参議院議員の名前を入力しても、その名前をもつ人物の情報であれば、人物が参議院議員でなくてもよい。

参議院議員50名を用いた結果を表4.7、表4.8に示す。50名のうち19名からプロフィールを収集した。ただし、19名のうち5名は不正解のプロフィールしか存在しなかった。したがって、正解プロフィールを収集していた人数は、全体の4分の1くらいであった。収集したプロフィールは、全部で31件あり、そのうち正解は23件であった。

著述家128名を用いた結果を表4.9、表4.10に示す。128名のうち109名からプロフィールを収集した。ただし、109名のうち5名は不正解のプロフィールしか存在しなかった。したがって、正解プロフィールを収集していた人数は、全体の81%であった。収集したプロフィールは、全部で588件あり、そのうち正解は459件（78%）であった。

プロフィールとして不適切なものを抽出した原因は、2つあった。1つは、前項で述べた原因と同じである。もう1つは、見出しに人名が存在するかどうかの判定方法にある。本システムは、人名の文字列が含まれていないかどうかで判定している。そのため、含まれていた文字列が人名そのものでなくても、見出しに人名が存在すると判定してしまう。例えば、図4.11に示すものがプロフィールとして抽出されていた。

表 4.7 実験 2 の人物情報の収集状況（参議院議員 50 名）

収集状況	正解プロフィールを含む	不正解プロフィールのみ	合計
プロフィールあり	14名	5名	19名
他の情報のみ	-	-	29名
情報なし	-	-	2名

個人情報リンクのみが9名。残り20名は、人物紹介ページのURLが存在した。

表 4.8 実験 2 の収集プロフィールの状況（参議院議員 50 名）

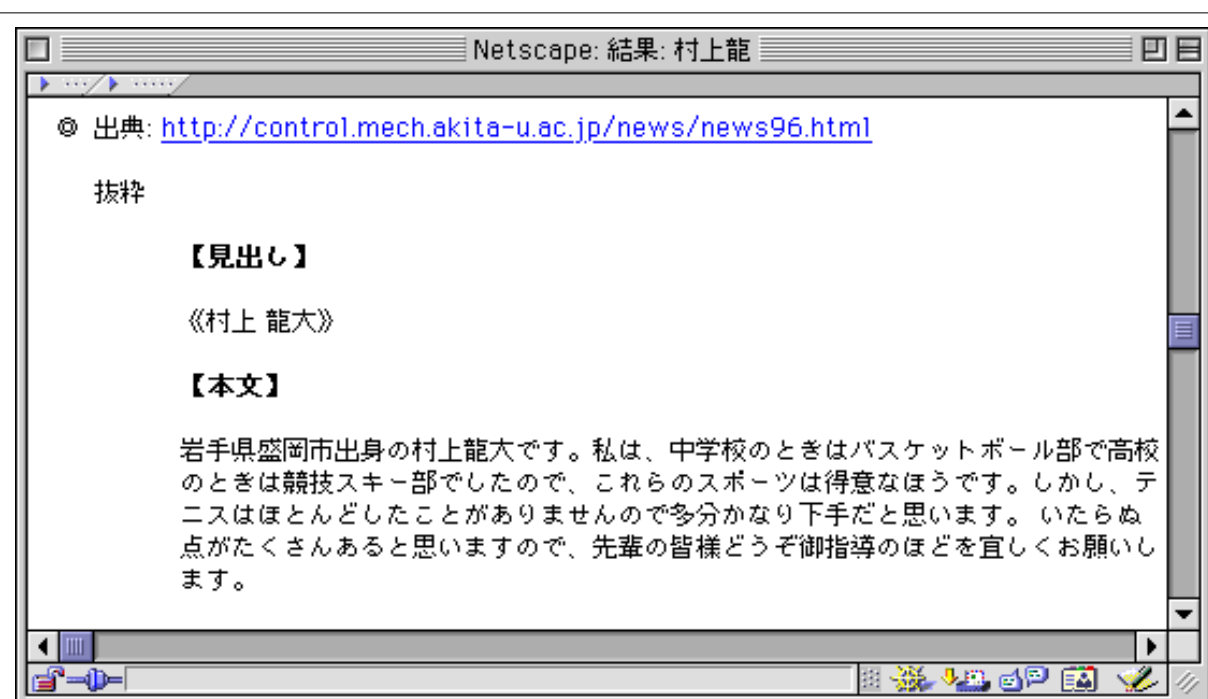
不要な部分	正解プロフィール	不正解プロフィール
ない	22件	-
ある	1件	-
合計	23件	8件

表 4.9 実験 2 の人物情報の収集状況（著述家 128 名）

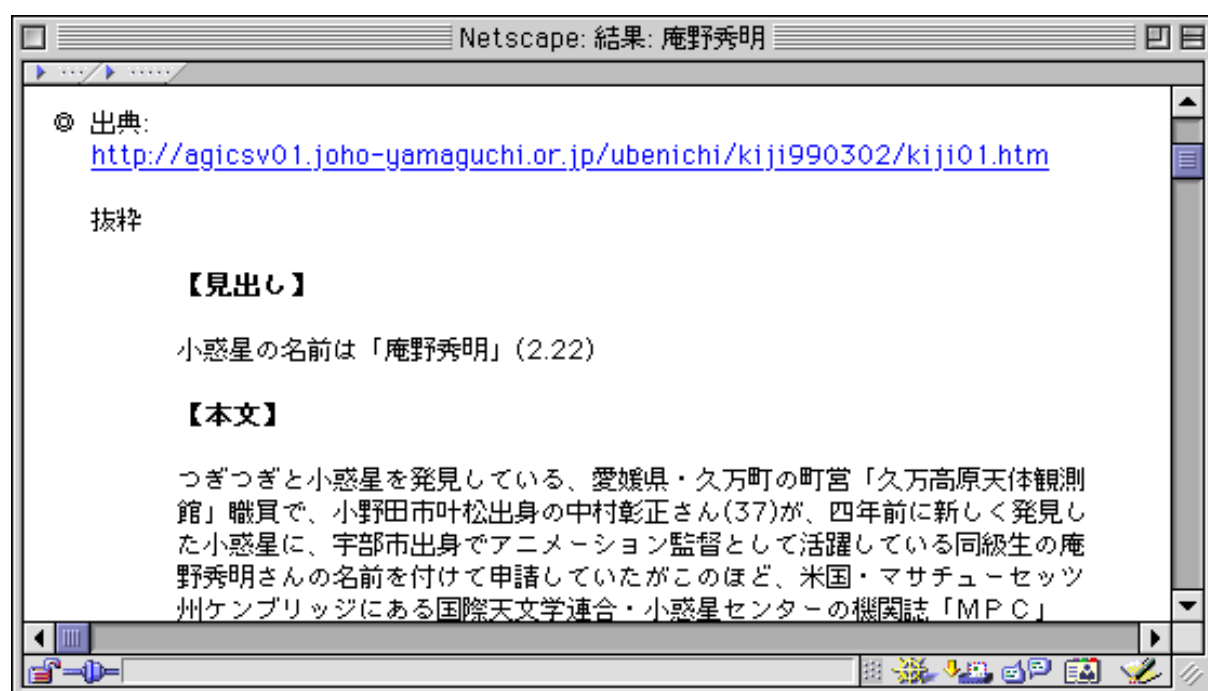
収集状況	正解プロフィールを含む	不正解プロフィールのみ	合計
プロフィールあり	104名	5名	109名
個人情報リンクのみ	-	-	5名
情報なし	-	-	14名

表 4.10 実験 2 の収集プロフィールの状況（著述家 128 名）

不要な部分	正解プロフィール	不正解プロフィール
ない	416件	-
ある	43件	-
合計	459件	129件



(「村上龍」を入力として、「村上龍大」の情報も収集している。)



(「庵野秀明」を入力として、小惑星の「庵野秀明」の情報も収集している。)

図 4.11 不正解プロフィールの例

4.4.3 検討

以下に，実験結果をまとめる．

- ・収集したプロフィールは，全部で 691 件あり，そのうち正解は 549 件（79%）であった．
- ・実験に用いた 178 名のうち 161 名（90%）から有用な情報を収集していた．ただし，参議院議員名を用いた場合は，プロフィールが収集されることは少なく，人物紹介ページの URL が収集されることが多かった．

本システムは，ウェブ上にプロフィールが存在する人物の名前が入力された場合，有用な情報を提供することが可能である．ただし，参議院議員のように，ウェブ上にプロフィールが多く存在しない人物に対しては，人物紹介ページの URL を 1 つだけ提供するという状況になることが多い．より多くの情報を提供するためには，ウェブ上にすでに存在しているプロフィールを抽出するだけでは不十分である．

第 5 章

検討と関連研究

本章では ,作成した人物検索システムの有効性について検討し ,関連研究について述べる .

5.1 検討

人物検索システムを評価する際に ,どのような人物に関する情報をどれくらい提供できるのか ,そしてその情報は信用できるのかが重要になる .これらについて ,以下に述べる .

(1) 情報提供できる人物

本システムは ,表形式の職業別人名録に掲載されている人物 ,および ,ウェブ上にプロフィールが存在する人物 ,の情報を提供することが可能である .ただし ,前者の人物に関しては ,収集できない人物も多く存在する .より多くの人物情報を収集するためには ,50 音以外で分割された人名録も収集するなど人名録の収集方法を改良する必要がある .また ,より多くの人物に対して情報提供を可能にするためには ,ウェブ上にすでに存在しているプロフィールを抽出するだけでは不十分である .

(2) 提供する人物情報

本システムは ,人物情報が掲載されたページの URL だけでなく ,表解析を用いて抽出した属性-属性値集合の人物情報 ,および ,レイアウト解析を用いて抽出したテキスト形式の人物情報 ,も提供する .実験の結果より ,これらの抽出情報は ,適切な

ものである可能性が高いことが明らかになった。したがって、利用者は、本システムの検索結果のページから有用な情報を効率良く収集することが可能である。なお、本システムは、収集した人物情報の統合を行っていない。利用者に提供する際には、抽出した人物情報と、その出典となるウェブページのURL、を対にしたものを列挙している。

(3) 情報の信頼性

本システムは、ワールドワイドウェブ上の情報を利用しているため、情報の信頼性の問題は非常に重要であるが、この問題の積極的な解決を行っていない。利用者自身が信頼できる情報であるかどうかを判断できるように、情報源のURLを提供することでこの問題を回避している。

5.2 関連研究

検索対象を人物とした既存のシステムには、個人のホームページを探し出すAhoy![1]や、gooのように人名からその人物に関するページを検索する機能をもったロボット型検索エンジンが存在する。これらは、URLを提供するのに対し、本システムは、URLだけでなく、ページに掲載されている人物情報の提供も行う点に大きな違いがある。

本システムの中心技術は、情報源となるウェブページの収集と、ウェブページからの情報抽出である。以下に、これらの技術の関連研究について述べる。

あるカテゴリのウェブページを収集する方法として、そのカテゴリ特有のURLパターンを利用する方法[1][2]や、そのカテゴリのリンク集を利用する方法[2][3][4]、そのカテゴリのホームページを利用する方法[5][6]が提案されている。本研究でも、これらが有効に働くことを利用する。

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利用できる[7]。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[8][9]。本研究では、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。

情報抽出の1つの方法として、表解析を用いる。ウェブ上には、表形式の人名録が多

数存在し、そのレイアウトは様々である。本研究では、このような様々な表の構造を正しく把握する必要があり、単純な表を解析する方法[10]では、不十分である。

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[11]がある。本研究では、ワールドワイドウェブ上に存在する職業別人名録を利用するという点に大きな違いがある。

第 6 章

結論

検索対象のカテゴリを限定した検索システムは、絞り込まれた結果を提供することができる。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情報を提供する人物検索システムを実現した。本システムにより、利用者は、効率良く人物情報を収集することができる。本システムが提供する人物情報は、次に示す 2 つの方法を用いて収集したものである。

第 1 の方法は、職名から人物情報を自動収集する方法である。まず、職名からその職業の人名録ページを自動収集する。次に、収集したページから表解析を行うことで属性-属性値集合で表現された人物情報を収集する。この方法を用いて行った実験結果は、次の通りである。

- ・「政治家」を対象とした実験では、2245 人の情報を収集した。ただし、「著述家」を対象とした実験では、833 人しか情報収集しなかった。
- ・「政治家」の場合は、すべての人物に対して適切な属性-属性値対応の情報を収集した。ただし、「著述家」の場合は、68 人に対して、1 つだけ不適切な対応の情報が存在していた。

これらの結果より、人物情報の内容がよいものを収集できることが明らかになった。ただし、職業によって、あまり多くの人物を収集できない。より多くの情報を収集するために、職業別人名録ページの収集に改良が必要である。

第 2 の方法は、人名から人物情報を自動収集する方法である。まず、人名からその人

物のプロフィールが存在するページを自動収集する .次に ,収集したページからレイアウト解析を行うことでその人物のプロフィールを収集する .この方法を用いて行った実験結果は ,次の通りである .

- ・ 実験に用いた178名のうち161名(90%)から有用な情報を収集していた .ただし ,参議院議員名を用いた場合は ,プロフィールが収集されることは少なく ,人物紹介ページのURLが収集されることが多かった .
- ・ 収集したプロフィールは ,全部で691件あり ,そのうち正解は549件(79%)であった .

これらの結果より ,ウェブ上にプロフィールが存在する人物の名前が入力された場合 ,有用な情報を収集できることが明らかになった .ただし ,参議院議員のように ,ウェブ上にプロフィールが多く存在しない人物に対しては ,有用な情報源が数URLしかないという状況になることが多い .より多くの情報を提供するためには ,ウェブ上にすでに存在しているプロフィールを抽出するだけでは不十分である .

謝辞

本研究を進めるにあたり , 佐藤理史助教授には , 多くの御教示を賜りました . 心より感謝致します . そして , 日頃から様々な御助言を頂いた佐藤研究室の皆様に感謝の意を表します .

参考文献

- [1] Jonathan Shakes, Marc Langheinrich and Oren Etzioni. Dynamic Reference Sifting: a Case Study in the Homepage Domain. WWW6, pp.189-200, 1997.
- [2] 大槻洋輔, 佐藤理史. ワールドワイドウェブを知識源とした地域情報の自動編集. 信学技報, Vol.99, No.534, pp.85-94, 2000.
- [3] 来住伸子. 分野を特定した自動収集による WWW 情報検索. 情処研報, NL124-12, pp.87-94, 1998.
- [4] Satoshi Sato and Madoka Sato. Automatic Generation of Web Directories for Specific Categories. Proc. of International Symposium on Digital Libraries 1999 (ISDL '99), pp.127-134, 1999.
- [5] Robert B. Doorenbos, Oren Etzioni, and, Daniel S. Weld. A Scalable Comparison-Shopping Agent for the World Wide Web. In Proceedings of the First International Conference on Autonomous Agents, 1997.
- [6] 有賀忠則, 佐藤理史. WWW を用いたカテゴリ別製品一覧表の自動編集. 情報処理学会第 60 回全国大会, 2000.
- [7] Dan DiPasquo. Using HTML Formatting to Aid in Natural Language Processing on the World Wide Web. Senior Honors Thesis, School of Computer Science, CMU, 1998.
- [8] Dayne Freitag. Information Extraction From HTML: Application of a General Learning Approach. Proceedings of the 15th National Conference on Artificial Intelligence, AAAI, 1998.
- [9] 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.

- [10] 伊藤史朗 , 大谷紀子 , 上田隆也 , 池田裕治 . 属性オントロジーの抽出と統合を用いた実空間と情報空間のナビゲーションシステム. 人工知能学会誌, Vol.4, No.6, pp.1001-1008, 1999.
- [11] 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.
- [12] 黒橋禎夫 , 長尾眞 . 日本語形態素解析システム JUMAN Version 3.5 使用説明書 , 京都大学大学院工学研究科 1998.

発表論文

- (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) 山本あゆみ, 佐藤理史. ワールドワイドウェブからの人物情報の自動収集. 信学技報, Vol.99, No.534, pp.93-100, 2000.

付録 A 実験結果

国会議員のサブカテゴリとなる役職名別に調べた結果を表 A.1 に示す．表内の「(国会議員)」には，役職名が国会議員であると判定したが，そのサブカテゴリまでは判定できなかった人物情報が該当する．以下の表も同様である．

都道府県議員のサブカテゴリとなる役職名別に調べた結果を表 A.2 に示す．

市町村区議員のサブカテゴリとなる役職名別に調べた結果を表 A.3 に示す．

表 A.1 国会議員の収集状況

役職名	URL数	ページ数	議員定数との比較結果
1.参議院議員	17	13	定数以上
2.衆議院議員	32	23	定数以上
(国会議員)	1	1	-
合計	50	37	-

表 A.2 都道府県議員の収集状況

役職名	URL数	ページ数	議員定数との比較結果
1.香川県議員	4	2	一致
2.徳島県議員	2	1	一致
3.長野県議員	1	1	一致
4.岐阜県議員	2	1	一致
5.奈良県議員	1	1	一致
6.栃木県議員	8	8	定数以下
7.京都府議員	1	1	一致
8.北海道議員	1	1	一致
9.岡山県議員	3	3	一致
10.埼玉県議員	2	2	一致
11.広島県議員	1	1	一致
12.岩手県議員	1	1	一致
13.群馬県議員	1	1	一致
(都道府県議員)	2	2	-
合計	30	26	-

表 A.3 市町村議員の収集状況

役職名	URL数	ページ数	議員定数との比較結果
1.文京区議員	1	1	一致
2.長岡京市議員	1	1	一致
3.名瀬市議員	2	1	一致
4.草加市議員	1	1	一致
5.神奈川県fujisawa_議員	1	1	一致
6.津久見市議員	1	1	一致
7.新区議員	1	1	一致
8.富士市議員	1	1	一致
9.徳島市議員	1	1	一致
10.静岡県shizuoka_議員	1	1	一致
11.上福岡市議員	1	1	一致
12.千葉県ichikawa_議員	1	1	一致
13.秋田県akita_議員	1	1	一致
14.大府市議員	1	1	一致
15.下関市議員	1	1	一致
16.黒部市議員	1	1	一致
17.滋賀町議員	1	1	一致
18.井川町議員	1	1	一致
19.西伊豆町議員	1	1	一致
20.甲府市議員	1	1	一致
21.大東町議員	1	1	一致
(市町村区議員)	2	2	-
合計	24	23	-

ローマ字は、URLから抽出したものをそのまま使用しているためである。