WWW上の職業別人名リストを利用した人名の収集

4P－8

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

1. はじめに

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．このような検索サービスを実現する1つの方法は，対象カテゴリに関するデータベースをあらかじめ作成しておき，これを検索時に利用する方法である．本稿では，人物を検索対象とした検索サービスの実現のために必要な人物情報データベースをWWWから自動生成する方法を提案する．

2. 人物情報自動収集システムの概要

作成したシステムの概要を図1に示す．本システムは，「政治家」などの職業名を入力とし，WWW上に存在する職業別人名リストを利用して，その職業をもつ人名と職業のサブカテゴリを収集し，人物情報データベースに格納する．人物情報の収集手順は，(1)人名リストページの収集，(2)職業のサブカテゴリ判定，(3)人名抽出の3ステップからなる．

2.1 人名リストページの収集

入力である職業名からその職業の人名リストが存在するページ(人名リストページ)を収集する．収集する人名リストの例を図2に示す．収集は，(1)，(2)の手順で行う．

(1) 人名リストページの候補の収集

人名リストページの候補となるページを次の2つの方法で収集する．

(a) 検索エンジンを用いた収集

職業名を表す言葉(「議員」)とリストを表す言葉(「名簿」，「一覧」，「紹介」)を組み合わせたもの(「議員名簿」など)が存在するページを検索エンジンを用いて収集する．

(b) リンク情報を利用した収集

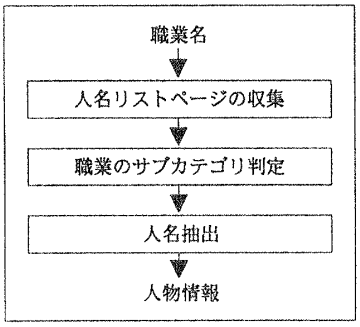

図1　システム構成

黒部市議会議員名簿

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

図2　人名リストの例

(a)で収集されたページ上に存在するリンクのうち，次に示す条件を満たす場合は，リンク先ページも合わせて収集し，さらに，それらのページに対して同様の処理を行う．これを最大3回繰り返す．

・アンカ文字列が職業名を表す言葉と一致する．

・アンカ文字列またはそれに併記されているリンク先ページに関する説明文が，職業名を表す言葉を含み，その後にリストを表す言葉または「検索」や「リンク」のような言葉が続いている(例えば，「県議会議員の紹介」，「政治家に関するリンク集」など)．

(2) 人名リストの有無判定

(1)で得られた候補ページに対して，(a)，(b)の処理により人名リストが存在するかどうかを判定する．

(a) 候補ページ内に対する人名リストの有無判定

次のいずれかに該当する場合，候補ページ内に人名リストがあると判定する．

・ページタイトルまたは見出しが，職業名を表

Collection of People's Names and Occupations from the World Wide Web.

YAMAMOTO Ayumi and SATO Satoshi.
School of Information Science, Japan Advanced Institute of Science and Technology.
Tatsunokuchi, Nomi, Ishikawa, 923-1211, JAPAN.

す言葉を含み，その後にリストを表す言葉が続いている．

・同一ページ内へリンクしている50音順の文字列がある

・50音順の文字列がある(アンカ文字列を除く)

ここでの50音順の文字列とは，「あいうえお順」，「50音順」，「あ行」，「あ」，「あ～お」のような言葉を指す．

(b) リンク先ページ内に対する人名リストの有無判定

候補ページ内のアンカが50音順の文字列であり，次の条件のいずれかを満たす場合，リンク先ページ内に人名リストがあると判定する．

・候補ページ内に人名リストがある

・50音順のアンカ文字列の一段上位の見出しに職業名を表す言葉がある

ここでの一段上位の見出しとは，何の「あ行」なのかを表す言葉を指す．

2.2 職業のサブカテゴリ判定

本システムは，単に与えられた職業をもつ人名を収集するだけではなく，それぞれの人名に対して，その職業のサブカテゴリを抽出する．ここで，サブカテゴリとは，例えば「政治家」の場合は，「衆議院議員」，「埼玉県議員」などである．

職業のサブカテゴリの判定は，下に示す4つの文字列の中に，あらかじめ用意されたサブカテゴリを表す言葉が存在するかどうかによって判定する．

・リストの見出し

・50音順のアンカ文字列の一段上位の見出し

・ページタイトルまたは見出し

・逆リンクのアンカ文字列とそれに併記されているリンク先ページに関する説明文

なお，職業名が政治家の場合は，この方法で判定できなければ，URLから都道府県または市町村区が特定できるかどうかを調べ，特定できる場合は，サブカテゴリをその都道府県または市町村区の議員とする．

2.3 人名抽出

サブカテゴリが判定された人名リストから人名を(1)から(3)の手順で抽出し，サブカテゴリと共にデータベースに格納する．

(1) リストの抽出

テーブルタグを用いたリストを対象とし，罫線が引いてあるテーブルと入れ子のないテーブルを抽出する．

表1　国会議員の結果

衆議院議員数(定数)	507 (500)
参議院議員数(定数)	253 (252)

表2　都道府県議員の結果

収集した議員リスト数	15（1都 1道 1府 12県※）
議員名を抽出したリスト数	10（1道 1府 8県）

※1県は一部の議員リスト

(2) リストの方向の判定

抽出されたリストに対して，縦，横のどちらの方向で見るのかをあらかじめ用意した属性名により判定する．属性名として，一般的な「氏名」，「生年月日」などと職業特有の「選挙区」，「会派」などを用いる．

(3) 人名の抽出

(2)で判定された方向で，氏名を表す言葉(「議員名」，「政治家名」，「名前」など)がある列または行を探し，その列または行の値を人名として抽出する．

3. 予備実験

「政治家」を入力とし，国会議員名と都道府県議員名を収集する実験を行った．その結果を表1，表2に示す．抽出した議員名に誤りはなかったが，国会議員は議員定数と一致しなかった．その原因として，字体の違いや旧リストと新リストの混合が挙げられる．また，都道府県議員名は，15都道府県の議員リストが収集でき，そのうち，10都道府県から議員名を全て抽出できた．

4. おわりに

本稿では，職業別人名リストの探索と人物情報の抽出方法について述べ，職業名が「政治家」の場合の実験結果について述べた．テキストから人物情報を抽出する研究として，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[1]がある．これに対し，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．今後は，「政治家」以外の他の職業に対しても同様の実験を行い，本システムの有効性を検証する予定である．

参考文献

[1] 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)