Web からの人物事典生成のための経歴情報の自動収集

Automatic Collection of Personal Histories for Generating Who's Who from the Web

木村 昌[1]　小山 聡[2]
田中 克己[3]

Rui KIMURA　Satoshi OYAMA
Katsumi TANAKA

Web 上には人物に関する情報が多く存在するが，Web を用いてユーザがある人物に関する情報を知るためには，多くのページを閲覧する必要があり，さらにページの中から所望の人物に関する記述を見つけ出す必要がある．我々の研究では，Web 上から人物情報を自動的に収集し，人物事典を自動生成する事で，ユーザが効率良く人物情報を手に入れる事を目的とする．

本稿では，人物事典の自動生成の第一段階として，ある特定の人物に関する Web 上の文書の集合から，西暦や元号など多様な表記をされる時間の表現を収集し，時間の表記法を統一する手法を述べる，またその時間表現と共に記述されている人物に関する記述を収集し，人物に関する記述を時系列順に提示する事で年表を生成する手法を提案する．

There is much information about people in the Web, but in order to know information about a specific person from the Web, users have to browse many web pages and they also have to find out statements about the person in each page. In our research, we make it our aim to automatically collect information about people from the Web, to automatically generate a Who's Who from the Web, and to help users know such information efficiently. In this paper, as the first stage of automatically generating a Who's Who from the Web, we propose a method to collect terms for time, which are written in various formats like western calendar or Japanese traditional era name, and to standardize a notation of these terms. In addition, we suggest a method to collect statements about people which co-occur with those terms. Using both standardized terms for time and statements about people, we present an automatically generated chronologic table of people.

1. はじめに

多様な情報がWeb上に存在する現在，知りたい情報を，Web検索エンジンを用いて調べるという事はもはや一般的なこととなってきている．

人物の情報に関しても店や商品の情報と同じように，Web検索をすることで情報を得ている．しかし，店や商品と異なり，人物の情報に関しては，情報がデータベースの形でまとまって掲載されているWebページは少ない．

このため，ある人物がこれまでどのような事を行ってきたかという情報を調べたい場合，多くのWebページを閲覧し，ユーザ自身がページ内から知りたい情報を見つけ出す必要があり，ユーザにとって手間となっている．また，人物の情報については伝記や年表といった時系列順に書かれた形式で情報を得る事に慣れ親しんでいるが，多くのページを閲覧する方法では，時系列に沿って情報を得ることができず，人物の全体像を把握することが難しい．

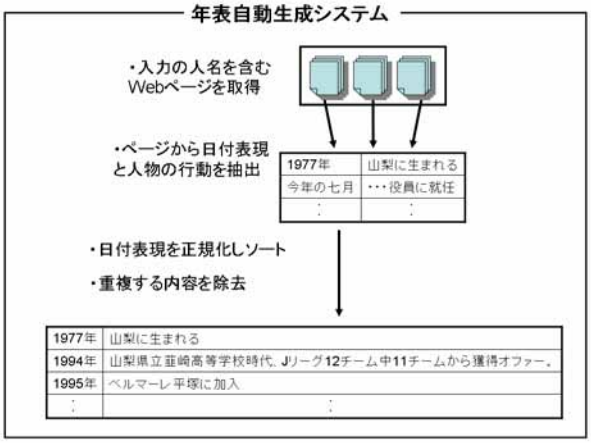

図1　システムの全体像

Fig.1　Overview of the system

我々は，このようなユーザの手間を解消し，Web上の知識を集約し提示することで，ユーザが目的の人物の全体像を容易に理解することができるように支援するシステムを構築することを目的とし，研究を進めている．

システムの全体像を図1に示す．ユーザが知りたい人物の名前を入力する．ページ内にその名前を含むページをWeb上から検索し取得する．取得した各々のWebページから，目的の人物の行動を収集する．それぞれの行動は「いつ」行われた行動かが分かる必要があるため，ページ中に現れる日付や年代を表す表現に注目し，人物の行動と共に，いつそれが行われたかを示す日付表現を抽出する．抽出された日付表現と行動を示す文を元に，同じ行動を示す文を集約し，重複する内容を省く．これら日付表現と人物情報のペアを用いて出力である年表を自動生成する．

出力としては，まず年表を生成することを想定している．これは，人物の説明に当たる文を変形することなく，日付表現を利用して時系列順に人物情報をソートすることで生成可能である．生成された年表を利用し，世の中一般の流れや，その人物に関係するような出来事をまとめ，時系列順で比較できるような形にする事で，人物の理解が深まるであろう．また，年表に掲載される情報から伝記や紹介番組などの他のコンテンツを生成することもできる．

システムの全体像には記述していないが，人名では同姓同名の人物が存在するため，複数の人物の情報が混在してしまう可能性もあるので，それを解決する必要が出てくることも

[1] 学生会員　京都大学大学院情報学研究科修士課程
kimura@dl.kuis.kyoto-u.ac.jp
[2] 正会員　京都大学大学院情報学研究科助手
oyama@dl.kuis.kyoto-u.ac.jp
[3] 正会員　京都大学大学院情報学研究科教授
tanaka@dl.kuis.kyoto-u.ac.jp

ある．筆者ら[1]は，Webページに含まれる同姓同名の人物を判別する研究を行ってきた．この研究では，検索エンジンによって出力される検索結果ページのみを利用し，個々のページの要約文であるスニペット中に含まれる単語を用いることで，ページを実世界の個人ごとにグループ分けしている．この手法を用いることで，あらかじめWebページを取得する段階で同姓同名の人物を同定しておく事ができる．

本稿では，人物事典生成の第一段階として，Webページから日付表現を抽出する手法と，その日付表現に付随する人物の行動を表す文を抽出する手法について述べる．

第2章では，人物情報の抽出や日付表現の抽出に関する関連研究を紹介する．第3章では，日付表現の抽出手法について述べる．第4章では，第3章で取得した日付表現に付随する，人物情報を収集する手法について述べ，第5章で本稿のまとめと今後の課題を述べる．

2. 関連研究

Webを情報源とした人物情報の抽出に関する研究は多く行われている．

KimらのArtequakt[2]では，Webから自動収集した情報を元に芸術家の伝記の自動生成を行っている．Artequaktでは，オントロジーと予め記述したパターンを元に情報を抽出している．Schiffmanら[3]は，ニュース記事の集合から人物の伝記を自動的に生成する手法を提案している．同格関係により人物の職業を導き出し，その職業と強い関連のある動詞が含まれる文を抽出している．

また，森ら[4]は，研究者の人名を対象にして，Web上の語の共起ネットワークを用いることで，Web上で人名と共起するキーワードを検出し，その研究者の所属組織・研究テーマ・共著者・プロジェクト名などの情報を抽出している．山本ら[5]は，職業別人名リストを利用し、表形式の職業別人名リストから表解析を行うことで人物情報を収集する手法と、列挙形式の職業別人名リストから人物プロフィールを抽出するシステムを提案している．

金田[6]は，百科事典から年代情報を抽出することで，ユーザが入力したキーワードに関する年表を動的に生成する手法を提案している．この研究では，年代表記と検索語とが近接して出現する箇所を検索し，年代順にソートすることで年表を生成している．

Webページ中の同姓同名人物の判別については様々な研究が行われている．WanらのWebHawk[7]では，人名によるWeb検索結果を同一人物ごとにクラスタリングし，それぞれの人を特徴付けるような単語を提示する事で，ユーザが所望の人物を人物の特徴から選択できるようにしている．白砂ら[8]は，Webページ内の文書構造やページのURL，抽出したプロファイル情報を利用して，人物の判別を行っている．

3. 日付表現の抽出

時系列の情報を収集するためには，まず，それぞれの情報がいつの情報であるかを示すような日付表現を抽出する必要がある．今回は人物事典の生成を行う目的で人物の一生に渡る情報を取得するため，抽出する時間情報の最低単位を日とし，抽出を行うこととする．

日付表現の抽出は，正規表現を用いて日付表現をヒューリスティックに記述する事で抽出を行った．表１に抽出対象とした日付表現の一部を示す．

表１に示した日付表現以外にも，「春」などの季節や，「数年後」や「下旬」などの曖昧な表現などが存在し，これらの言葉によって時期を表す事があるが，今回は年月日を特定する事ができないという理由で，抽出対象からは除いている．

表１　抽出対象とする日付表現の例

Table 1　Examples of date expressions to be collected

日付表現のタイプ	例
数字と共に「年」「月」「日」が明記されている	2005 年 6 月 8 日 二千五年三月
「/」や「-」などで年月日が区切られている	2005/4/13 2006-05-12
数字・「年」と共に和暦の年号を含む	平成 18 年 昭和五十七年
時間経過を指し示す	翌年
ある時点を指し示す	その年の大晦日
ある特定の期間を曖昧に示す	年末

本研究では正規表現を用いてWebページから日付表現の抽出を行った．抽出の対象とするWebページは，人名を質問キーワードとして用いてWeb検索エンジンGoogleで検索を行い，検索結果の上位100件のページを対象とした．また，質問とした人名は，Webページに掲載されている情報が比較的多い，研究者，スポーツ選手，芸能人，政治家といった職業から５種類の人名を選択した．

また，年・月・日のうち，連結可能なものは全て連結されており，連結されたものは一つの日付表現として抽出している．例えば，「2005年」と「年末」がスペースや格助詞「の」を挟み，隣り合って出現している場合，「2005年　年末」「2005年の年末」といったように連結する事で，一つの日付表現とみなした．

3.1　日付表現の意味補完

日付表現を抽出した後，自然言語で書かれている各々の日付表現が意味する年月日を全て正規化する必要がある．今回は，年月日をそれぞれ4桁，2桁，2桁の数値で表し，8桁の数値に変換する事で正規化を行った．例えば，「1982/5/3」は19820503に変換される．

また，「翌年」のようにある時点からの時間経過を指し示す日付表現など，その語だけでは年月日の正規化を行うことができない表現が存在する．人間がこのような表現を見た場合，周りの文脈などから日付表現の意味を補完し，その語が示す時点を推定する事ができる．本稿では，人間が行う日付表現の意味補完を，補完対象の直前に出現する日付表現の情報で意味補完することで近似するルールを設定し，意味補完の精度や問題点を調査するため実験を行った．以下に示すのは，今回利用した意味補完のためのルールの一部である．

- 「五日」のように月（年）を表す表現がない場合，直前の日付表現が意味する月（年）で補完
- 「今月」や「同年」などの時間遷移がないと判断できる表現の場合，直前の日付表現が意味する月（年）で補完
- 「昨日」や「５ヵ月後」や「半年前」など，時間遷移を意味する日付表現の場合，直前の日付表現から時間遷移させ，該当の日付表現が意味する時点を計算し，補完

表2はWebページから抽出できた日付表現を分類し，補完対象となった日付表現数を人名ごとに表にまとめたものである．平均すると，1ページに約105個の日付表現が存在し，そのうち約27個は意味補完が必要な日付表現となっている．つまり3/4の日付表現を用いて1/4の日付表現の意味補完を行うこととなる．また，直前の日付表現に年月日を含まない

場合，ルールの適用が不可能であると定義した．

表 2　抽出された日付表現
Table 2　Collected date expressions

人名	日付表現数	補完が必要	ルールの適用が不可能
田中克己	932	216	64
小山聡	777	233	56
小泉純一郎	1296	287	78
中田英寿	1308	345	82
蛯原友里	951	251	45
合計	5264	1332	325

実際に 1334 個の日付表現を上記のルールを用いることで意味の補完を行い，それぞれを人手で正解か不正解かを分類した．また，“補完が不可能な表現”については，人がページの URL やタイトルを含む Web ページのテキストのみで日付表現が指し示す時点が分かる場合は“人は解決可能”，そうでない場合は“人も解決不能”とした．例えば URL が「http://～/20050602/index.html」といった場合に，この記事は 2005 年 6 月 2 日に記述されたものであると人は理解できる．また、ページタイトルが「DBWS2006」といった場合， 2006 年の出来事であると推定できる．人手で補完対象の日付表現を判定した結果を表 3 にまとめた．

表 3　補完した日付表現の正否
Table 3　Results of completion

補完の結果	日付表現数
正解	595
不正解	412
人は解決可能	182
人も解決不能	143
合計	1332

補完対象となった日付表現のうち，正解となったものは約 59%であり，単純な意味補完の手法であっても比較的良い精度を得られた．また，ルールでは補完不可能とした日付表現については，約 61%のものは人が解決できるものであった．この事から，より複雑なルールや URL などテキスト以外の情報を利用することで，精度を伸ばすことができると考えている．

3.2　日付表現の抽出における課題

今回，抽出の目的とした日付表現は，年表の生成を目的とし，ページ内に記述されている人物の行動などがいつ行われたかを示すような日付表現である．しかし，抽出された日付表現の中には，目的とは異なる日付表現や実際には日付表現ではない言葉が誤って抽出されているものも多く，このようなノイズが意味補完の際の精度を下げてしまっていることが分かった．

これらの日付表現は，「三日月」「五日市」のように人名や地名など固有名詞の一部になっているものや，「六年ぶり」「1 日延長した」のように時点として抽出したが時間遷移を表しているもの，「一年間」のように期間を表しているもの，「昨年同様」などの行動が行われた時点を表さないものなど，いくつかのパターンに分類可能である．

これら全ての誤りに対し，正規表現を利用して取り除く事は難しい．期間や時間遷移を表すものに関してはパターンを記述する事で取り除くことは可能であるが，固有名詞に関しては，出現する全てのパターンを記述する事は現実的ではない．これについては，形態素解析を用い，日付表現であるか判定することで対応できる可能性があるので，今後検証していきたい．また，固有名詞以外のタイプについては，抽出の際の正規表現にパターンを追加し，意味補完のルールも追加することで，今後対応していきたい．

3.3　日付表現の意味補完における課題

意味補完の手法としては，直前の日付表現の情報のみを利用して意味補完を行うという最も簡単な方法で実験を行ったが，意味の補完が不可能な例や，誤って意味補完してしまう例が目立ち，今後解決すべき課題が見つかった．ここでは，目立った補完誤りや，その対処方法について検討する．

意味補完における誤りとしては，時間遷移を表すものや，「今年」といったある時点を指し示すものに誤りが多かった．ルールでは，時間遷移については直前の表現からの時間遷移として見ていたが，Blog 記事やニュース記事の場合，これらの日付表現は，記事の執筆時期から見た時間遷移を意味する表現がほとんどである．このため，多くの補完誤りを生んだ．特に有名人の人名は，Blog 記事やニュース記事に現れることが多く，この問題は解決する必要がある．

さらに，Blog サイトやニュースサイトの場合，メニュー，コメント，トラックバックなど，記事の文脈とは関係のない文章が多く含まれており，これらの部分で多くの補完誤りが生まれていた．

また，ショッピングサイトでは，「前日までの在庫」「次回入荷日は 6/3 です」といったようにアクセスするたびに指し示す日付が変わるもが多く，不正解となる場合が多かった．

これらに対処するためには，取得したページを Blog 記事，ニュース記事，ショッピングサイト，その他のサイトなどに自動分類し，ページの種類ごとに手法を変える必要があるだろう．Blog 記事・ニュース記事については記事本文のみを抽出対象として，それに加えて記事が書かれた日付を取得する必要がある．また，ショッピングサイトに関しては人物情報と共に出現するような日付表現自体がほとんど無いことから，今後はショッピングサイトと判断された場合はページを解析対象から省く事も検討する必要がある．

徳江ら[9]は，ニュース記事内に出現する年を表す日付表現を抽出し，ニュース記事の掲載日を参照し利用する事で意味解決を行っている．また，Blog 記事内の日付表現に関しては，南野ら[10]による研究内で日付表現の抽出が行われている．この研究では，繰り返し出現するタグのパターンなどから，Blog 記事の執筆された日付を抽出している．

今後はこれら関連研究の手法を検討し，各種 Web ページ内に出現する日付表現を抽出する必要があると考えている．

4. 日付表現に付随する人物情報の収集

4.1　人物情報の収集手法と実験

本研究で抽出を目指す文は人物の行動を表す文である．「2006 年 5 月　ワールドカップ直前のドイツ戦では、3-5-2 のボランチの位置でスタート」「2002 年よりファッション雑誌 CanCam の専属モデルを務める。」といった，日付表現の直後に人物の行動を表す文がある場合を収集対象とし，その手法について検討する．

今回提案する手法は以下の流れとなる．

1　第 3 章で抽出した日付表現に続く文を取得する
- 全ての HTML 文書が句点を含む文章ではないため，改行で終わるものも取得する

2　文の終わりが名詞の場合，サ変接続の名詞かどうかを判

定し，そうでなければ除外する
- サ変接続の名詞は「購入」のように，名詞で終わるが行動を示している

3 日付表現の直後の文字が「に」,「より」,「から」,「、」,空白となっているものだけ残し，それ以外は除外する
- 適合率が高いと思われるこのパターンのみを収集

4 収集した文が日本語を含むもののみ抽出する

この手法で，文を取得し，その文がどの程度適合しているかを調べた．質問キーワードが「中田英寿」の場合，1308の日付表現中，上記手法で収集された文は 207 文であった．そのうち，中田英寿の行動と関わりがあるとみなせるものは 166 文であり，適合率は 79.2%であった．また，この収集された 207 文のうち, 日付表現の抽出の実験を行った 2006/6/9 以前を意味する日付表現は 192 文であり，そのうち適合しているものは 164 文（適合率 85.4%）と高い適合率を示した．

4.2 人物情報の収集実験結果の考察

日付表現を抽出した日以降の話題, つまり未来の出来事に限っては 15 文中 2 文が適合していると判定されたように，人物情報の収集対象としては適していないことが分かった．

また，今回の実験では非常に簡単なルールを定めることで人物情報を示す文を収集してみたが，79.2%という高い適合率を得ることができた．不適合とした文の中には，Web ページ特有の言葉である「更新」といった言葉や，月ごとの占い結果を表示している文が多く目立ったため，これらを取り除くことにより，より高い適合率を得ることができるだろう．

今回の実験では再現率を測ることができなかったが，今後の実験では，我々の方で収集したい文章を予め定めておき，その文がどの程度の割合で収集できたかで再現率を測る必要がある．また，実験で用いた質問キーワードが人名一人分のみであったので，今後データ数を増やし，さらなる実験を進める必要がある．

5. まとめ

本稿では，人物事典の自動生成に向けて，研究の第一段階として，年表の自動生成に必要な日付表現と人物情報を定義し，それぞれを簡単なルールを定義することで Web ページから抽出する実験を行った．また，Web ページ内のテキストを利用することで，日付表現の意味を補完する実験を行った．

実験の考察において，適合率を高めるための多くの課題が見つかったが，今回は課題を解決する手法を提案するにとどまった．今後は，今回提案した手法を用いて実験を進め，年表の自動生成に向けて，精度の高い日付表現と人物の行動を抽出するシステムを実装していきたい．

[謝辞]

本研究の一部は，文部科学省科学研究費補助金若手研究(B)「参照の同一性判定に基づく複数 Web ページの検索閲覧方式の研究」（研究代表者：小山聡，課題番号：16700097），文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築 - 異メディア・アーカイブの横断的検索・統合ソフトウェア開発」（研究代表者：田中克己），および，文部科学省 21 世紀 COE 拠点形成プログラム「知識社会基盤構築のための情報学拠点形成」（リーダー：田中克己，平成 14～18 年度）によるものです．ここに記して謝意を表します．

[文献]

[1] 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(DEWS2006)論文集, 2C-i11, Mar. 2006.

[2] S. Kim, H. Alani, W. Hall, P. Lewis, D. Millard, N. Shadbolt, and M. Weal, "Artequakt: Generating Tailored Biographies from Automatically Annotated Fragments from the Web," In Proceedings of Workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM'02), the 15th European Conference on Artificial Intelligence, (ECAI'02), pp. 1-6, 2002.

[3] B. Schiffman, I. Mani, K. J. Concepcion, “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), July. 2001.

[4] 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.

[5] 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,” 情報処理学会研究報告, No. 2000-ICS-119, p.165-172, 2000.

[6] 金田泰, “百科事典から動的に年表を生成するテキスト検索法のための年代情報の抽出法と表現法,” 情報処理学会 情報学基礎研究会報告 Vol.1999, No.57, pp.81-88, 1999.

[7] X. Wan, J. Gao, M. Li and B. Ding, "Person Resolution in Person Search Results: WebHawk," In Proceedings of CIKM'05, pp. 163-170, 2005.

[8] 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(DEWS2006)論文集, 2C-i7, Mar. 2006.

[9] 徳江英範, 白井清昭, "対話型質問応答システムにおける質問の曖昧性の検出," 第 11 回言語処理学会年次大会, pp. 1092-1095, Mar. 2005.

[10] 南野朋之, 奥村学, "なんでも RSS - HTML 文書からの RSS Feed 自動生成," 人工知能学会第 10 回セマンティックウェブとオントロジー研究会(SIG-SWO-A501), 2005.

木村 塁 Rui KIMURA
京都大学大学院情報学研究科修士課程在学中．2005 年京都大学工学部情報学科卒業．主に Web データマイニングの研究に従事．日本データベース学会学生会員．

小山 聡 Satoshi OYAMA
京都大学大学院情報学研究科社会情報学専攻助手．2002 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．主に機械学習，データマイニング，情報検索の研究に従事．電子情報通信学会，情報処理学会，人工知能学会，日本データベース学会，IEEE，ACM，AAAI 各会員．

田中 克己 Katsumi TANAKA
京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了．京大工博．主にデータベース，マルチメディアコンテンツの処理の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，情報処理学会，日本データベース学会等各会員．