Webからの履歴書作成のためのイベント情報の抽出

Event Extraction for Creating Curriculum Vitae from the Web

上田 洋[*1] Hiroshi UEDA
村上 晴美[*2] Harumi MURAKAMI
辰巳 昭治[*1] Shoji TATSUMI

[*1] 大阪市立大学大学院工学研究科
Graduate School of Engineering, Osaka City University

[*2] 大阪市立大学大学院創造都市研究科
Graduate School for Creative Cities, Osaka City University

We propose an event extraction method for creating curriculum vitae from the Web. One advantage of creating curriculum vitae is that users can easily understand the designated person. In this research, an event is a sentence that includes both time and events related to a person. Our method is based on the following: (1) extracting events using heuristics, (2) filtering, and (3) judging events related to a person by mainly using the patterns of HTML tags. Experimental results revealed the usefulness of our proposed method.

1. はじめに

我々は，Web 上の同姓同名人物を分離して人物属性情報を表示するシステム[上田 07]を開発している．[上田 07]では，クラスタリング手法により Web ページを同姓同名人物毎に分離し，人物の識別を容易にするために，分離された Web ページのクラスタから人物のラベルとして人物属性情報（関連地方，関連職業，関連キーワード）を抽出し，提示している．しかし，[上田 07]の同姓同名人物の分離，提示する情報どちらも，精度に問題があった．そこで我々は，同姓同名人物の分離の精度向上[片岡 08]と，職業に関連する情報[上田 09]や位置情報[高守 09]の抽出を検討した．これらにより，同姓同名人物の分離性能が向上し，従来よりも各人物の識別が容易になったと考える．しかし，ラベルとして抽出する [上田 09，高守 09]はどちらも 1 つと判断するための情報が少ないため，各人物についてある程度の知識がないと識別できない可能性がある．また，あまり良く知らない人物については，これらの手法にて得られた情報を提示しても識別は困難である．そこで，Web ページから，より詳しい人物の理解に役立つ情報を抽出，整理する手法を検討する．

我々は，詳しい人物の理解に役立つ情報として，履歴書に記述される情報に着目した．履歴書は，人物の履歴を記した書類である．履歴書は，就職活動の選考資料としてよく用いられるため，多くの人が作成や閲覧の経験がある．そのため，多くの人になじみのある書類であると考える．人物の履歴は，人物の理解に役立つ情報の一つである．履歴書に記載される履歴の多くは，「2005 年 3 月 31 日 大阪市立大学 卒業」や，「昭和 53 年 9 月 1 日 豊橋技術科学大学赴任」のように，時間（年月日）と人物に関する出来事を含む．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」をイベント情報と呼び抽出対象とする．人物に対する「いつ」「何をしたか，何が起ったか」という情報を別々でなく，あわせて提示したほうが，単体で提示するよりも人物を理解する上で役に立つと考える．

本研究では，詳しい人物の理解に役立つ情報として Web ページから履歴書を作成することを目的とし，イベント情報を抽出する手法を提案する．

連絡先：上田 洋，大阪市立大学大学院工学研究科，
d06tb001@ex.media.osaka-cu.ac.jp

2. 提案手法

本研究では，Web での人名検索により得られた Web ページから人物と関係があるイベント情報を抽出する．まず，ヒューリスティックを用いて Web ページからイベント情報を取得する．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，人物とイベント情報が関係あるかどうか SVM を使用して判定する．

2.1 イベント情報の取得

Web ページからイベント情報を取得する．本研究では，HTML で記述された Web ページを対象とする．取得対象とするイベント情報は，年（例えば，2008 年，’08，平成 20 年など）が出現する文である．まず，ヒューリスティックにより Web ページから文を切り出し，得られた文に西暦か年号（本研究では和暦のみ対象とする）の形式で記述された年を含むかどうかを判定する．

文の取得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

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.で挙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．以下の形式を含むものをイベント情報として取得する．

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）

[1] 明治以降を対象とした．

2.2 フィルタリング

得られたイベント情報には，履歴書に掲載には適さないものが多く含まれる．そのため，あらかじめ選定した不要語と不要パターンを用いてイベント情報のフィルタリングを行う．

不要語として，21 種類を選定した．選定した不要語には，「許諾」「転載」「投稿」「レビュー」「copyright」などがある．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出現する．これらが出現する文のほとんどは，人物とは関係のない文である．

不要パターンとして，20 種類を選定した．不要パターンの多くは，「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含む文を対象としたものである．詳細な時間を含む文の多くは，Web ページやブログの更新時間，コメント，掲示板の発言の時間などである．これらには人物に関係のない内容がほとんどであり，仮に人物についての発言であっても，履歴書に掲載できる内容が含まれる可能性は低いと考える．

表 1　人物との関係性判定に用いるパターン

1. イベント情報に氏名を含む
2. イベント情報に姓を含む
3. イベント情報に含まれる名詞の数
4. イベント情報に含まれる対象人物以外の人名の数
5. タイトル内に検索氏名を含む
6. タイトル内に姓を含む
7. タイトル内の名詞の数
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. イベント情報の最も近くに出現するh1～5タグ内の氏名を含む
12. イベント情報の最も近くに出現するh1～5タグ内に姓を含む
13. イベント情報の最も近くに出現するh1～5タグ内の名詞の数
14. 最初のtrタグ内に氏名を含む*
15. 最初のtrタグ内に姓を含む*
16. 最初のtrタグ内の名詞の数*
17. イベント情報の出現する行の最初のtdタグ内に氏名を含む*
18. イベント情報の出現する行の最初のtdタグ内に姓を含む*
19. イベント情報の出現する行の最初のtdタグ内の名詞の数*
20. イベント情報とその前方の最も近くに出現する氏名との間の名詞の数
21. イベント情報とその前方の最も近くに出現する氏名との間の対象人物以外の人名の数
22. イベント情報とその前方の最も近くに出現する氏名との間の各タグの出現数
23. イベント情報とその前方の最も近くに出現する姓との間の名詞の数
24. イベント情報とその前方の最も近くに出現する姓との間の対象人物以外の人名の数
25. イベント情報とその前方の最も近くに出現する姓との間の各タグの出現数
26. イベント情報とその後方の最も近くに出現する氏名との間の名詞の数
27. イベント情報とその後方の最も近くに出現する氏名との間の対象人物以外の人名の数
28. イベント情報とその後方の最も近くに出現する氏名との間の各タグの出現数
29. イベント情報とその後方の最も近くに出現する姓との間の名詞の数
30. イベント情報とその後方の最も近くに出現する姓との間の対象人物以外の人名の数
31. イベント情報とその後方の最も近くに出現する姓との間の各タグの出現数

注：氏名と姓は対象人物のもの
*イベント情報がtableタグ内にある場合

2.3 人物との関係性判定

これまでの処理にて得られたイベント情報は，検索対象の人物の氏名を含む Web ページから取得している．そのため，人物と関係がある内容を含む可能性は高いが，人物と関係のない内容も含まれる可能性もある．そこで，人物と得られたイベント情報が関係あるかどうかを判定する．

HTML で記述された Web ページは，構造的に記述されたページがある一方，全く構造を意識せずに記述されたページもある．また，タグの閉じ忘れなどの HTML の記述に間違いのあるページもある．このように，HTML で記述された Web ページはかなり雑多であるため，従来のプレーンテキストを対象とした手法や DOM などの階層情報を用いる手法では，有効性に限界がある．プレーンテキストを対象とした手法は，文脈から該当人物との関係性を推定することが一般的だが，構造化された文書では文の形式をとらない記述も多くあり，文脈からの推定ができない場合が多い．DOM など階層情報を用いる場合，文書が構造化されていることが前提となる．上述のように，全く構造を意識せずに記述されている場合や，HTML の記述ミスがある場合，有効に機能しない可能性がある．

本研究では，全く構造を意識せずに記述されている場合や，HTML の記述ミスがある場合でも正常に判定できるように，人物の氏名または姓とイベント情報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．人物の氏名または姓とイベント情報の間の各タグの出現パターンであれば，階層構造となっていない場合や多少の HTML の記述ミスがある場合でもパターン生成は可能であり，高い汎化性能を持つ SVM により学習させることにより，従来手法では対応できないケースでも正常な判定が可能となると考える．

人物との関係性判定に用いるパターンを表 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について対象とするためにパターン化する．表 1 の 5 から 13 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パターンを取得している．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, 24, 25, 28, 31 については，人物の氏名または姓とイベント情報の間のタグの出現数をパターンとする．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．

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[2]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 Web ページに対し，2.1, 2.2 節の処理を行い，7211 のイベント情報を取得した．7211 のイベント情報に対し，人手により人物と関係があるかどうかを判定した．その結果，人物と関係があるイベント情報（正解データ）2266，人物と関係がないイベント情報（不正解データ）4474 に分離し[3]，教師データとした．教師データから，表 1 のパターンを生成し SVM により学習した．

学習の結果得られたデータを元に，未判定のイベント情報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．関係がないと判定されたイベント情報については以下の処理から除外する．

表 2　解析対象タグ

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

[2] 人物は職種に偏りがあるが，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．

[3] 人物と関係があるか判定不能のものは除外した．

3. 実験

提案手法の人物と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．

3.1 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（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 の氏名（以下，受賞者）を用いた（表 3 参照）．

26 の氏名を使用し，Google Web APIs を用いて Web ページを検索 した結果，得られた上位 50 件を処理対象とした．得られた Web ページについて，2.1 節，2.2 節の処理を行い，9004 件のイベント情報を得た(表 3 参照)．この 9004 件のイベント情報を，後述の実験にて使用した．

表 3 データセット

氏名	職業	イベント情報
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

3.2 方法

実験を行う前に，9004 件のイベント情報について各受賞者と関係があるイベント情報かどうかを人手で判定した．判定したデータを受賞者と関係があるイベント情報と関係がないイベント情報に分類し，正解データとした．

比較手法として，HTML の階層構造を用いた手法と，プレーンテキストを用いた手法を用いた．各手法の適合率と再現率，F 値を比較した．適合率，再現率，F 値は，以下のように定義する．

$$precision = \frac{N \cap R}{N}$$

$$recall = \frac{N \cap R}{R}$$

$$F - measure = \frac{2 \cdot precision \cdot recall}{presicion + recall}$$

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されたイベント情報の数，R は人手により受賞者と関係があると判定したイベント情報の数である．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．

HTML の階層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階層構造のパターンを生成する手法として k-照合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，階層構造を解析の対象とするタグについては，主に構造化のために使用される可能性の高いタグのみを対象とした．具体的には，2.3 節にてパターンに使用しているタグと同じもの を対象とした．対象のタグについて，k-照合部分木を用いてパターンを生成する．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 にて学習する．

プレーンテキストを用いた手法は，HTML の構造情報を用いず，記述された文章のみから判定する手法である．パターンとして，提案手法にて使用しているパターンのうち構造情報以外のパターンを用いた．具体的には，表 1 の 1 から 4 と 20，21，23，24，26，27，29，30 である．

なお，HTML の階層構造を用いた手法，プレーンテキストを用いた手法，どちらも教師データには提案手法と同じデータを用いた．

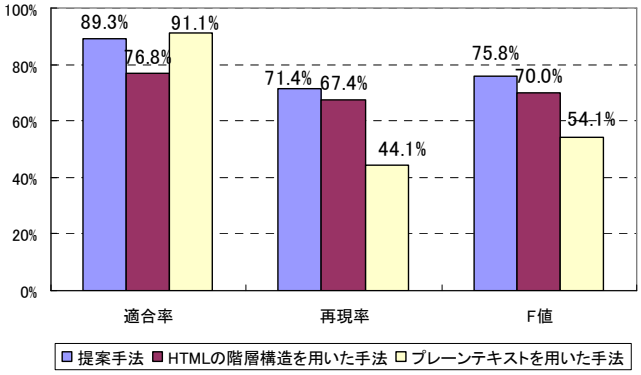

図 1 実験の結果

3.3 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，かつ，再現率は 3 手法のうち最も高い評価で，7 割を超える

正解のイベント情報を判定できた. 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 1 参照).

HTML の階層構造を用いた手法は, HTML 全体の論理構造を解析しなければならない. 論理構造の解析は, タグ構造が完全であることが前提となる. そのため, タグの記述ミスなどがある Web ページでは, 正常な論理構造が得られない場合がある. 実験でも, パターンの生成に失敗したイベント情報がいくつかあった. 提案手法では, 主に HTML のタグの出現パターンとして用いているため, HTML 全体の論理構造の解析は必要ない. 実験では, 全てのイベント情報でパターンの生成に成功した. また, 論理構造が複雑な HTML では解析に時間がかかる可能性がある. 提案手法では, HTML 全体の論理構造の解析を行わないため, 論理構造が複雑な HTML の場合, HTML の階層構造を用いた手法よりパターンの生成が早くなる可能性がある. HTML の階層構造を用いた手法のその他の問題として, 抽出されたイベント情報の周辺の文字列の情報が考慮されていない点がある. 提案手法では, イベント情報の周辺の文字列の情報も考慮している. そのため, HTML の構造情報を用いた手法よりも適合率が高くなったと考える.

プレーンテキストを用いた手法については, 受賞者と関係があると判定されたイベント情報の多くが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 の構造情報もパターンとして用いているため, 多くの受賞者の氏名または姓を含まないが受賞者と関係があるイベント情報を判定できた.

4. 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. [Kimura 07]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階層構造をパターンに判定を行っている. 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, 有効に機能しない場合がある. 本研究では, 階層構造ではなく, 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. [Kimura 07]では, 経歴を抽出の対象としている. 一般的な履歴書でも, 経歴が記述される. 本研究でも, 今後, 得られたイベント情報から経歴を選別する処理を検討したい. [Kim 02]ではニュース記事を, [Schiffman 01]では Web 上から得られた情報を用いて, 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, XML を対象に人物と文の関連性を用いて求めている. 本研究では, HTML を対象としている. 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. 本研究の人物との関係性判定でも, 表(table)を考慮している.

5. おわりに

本研究では, 詳しい人物の理解に役立つ情報として Web ページから履歴書を作成することを目的とし, イベント情報を抽出する手法を提案する. 本研究におけるイベント情報とは, 時間と, 人物に関する出来事の両方を含む文である.

提案手法は, ヒューリスティックを用いたイベント情報の抽出, 抽出により得られたイベント情報のフィルタリング, 主に HTML のタグの出現パターンを用いた人物との関係性の判定, から構成される.

人物と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. これらは提案手法の人物との関係性判定の有効性を示す結果であると考える.

一般的な履歴書は, 学歴, 職歴などのカテゴリ毎に記載される. 今後, 提案手法にて得られたイベント情報を学歴, 職歴などのカテゴリ毎に分類するなど, より履歴書に見える形式に提示できるようにイベント情報を整理する手法を検討したい.

参考文献

[上田 07] 上田 洋, 村上 晴美: Web 上の同姓同名人物を分離して人物属性情報を表示するシステム, 第 21 回人工知能学会全国大会論文集, 3G8-1(2007)

[片岡 08] 片岡 真一, 上田 洋, 村上 晴美, 辰巳 昭治: 人物名に着目した二段階クラスタリングによる Web 上の同姓同名人物の分離, 第 22 回人工知能学会全国大会論文集, 1E1-4(2009)

[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(2009)(採録決定)

[高守 09] 高守 雄也, 上田 洋, 村上 晴美: Web ページからの人物に関する位置情報の抽出, 第 71 回情報処理学会全国大会講演論文集(2009)

[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)

[Kimura 07] R. Kimura, S. Oyama, H. Toda, and K. Tanaka: Creating Personal Histories from the Web using Namesake Disambiguation and Event Extraction, In Proceedings of ICWE 2007, pp. 400-414 (2007)

[Kim 02] 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), pp. 1-6 (2002)

[Schiffman 01] B. Schiffman, I. Mani, and 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)(2001)

[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)

[Yoshida 04] M. Yoshida, K. Torisawa, and J. Tsujii: Integrating Tables on the World Wide Web, 人工知能学会論文誌, Vol. 19, No. 6. pp. 548-560(2004)