2D2-3

Web 上の同姓同名人物識別のための職業関連情報の抽出

Extracting Vocation-Related Information from Person Search Results to Identify Different People with Identical Names on 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 have been developing a system that classifies different people from person search results and displays a list of people by related prefectures, vocations, and keywords. Previously we extracted vocations using a list of vocations from Wikipedia, but the method's precision was not successful. In this paper, we propose a method to extract vocation-related information from the Web pages of person search results. Vocation-related information is more useful for users to identify a person. An example of vocation-related information is "professor of Engineering, Osaka City University," which is more useful than just a single word: "professor."

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，同姓同名人物の識別を容易にするために，人物のラベルとして人物属性情報を提示している．しかし，人物属性情報の一つとして表示している職業に関する情報の提示精度は必ずしも良くなかった．そこで本研究では，より精度の高い職業に関する情報（職業関連情報）を Web ページ内から抽出する手法を提案する．

2. 提案手法

人物毎に分けられた Web ページクラスタに職業関連情報(職業を表す語，所属と役職を表す語，著作と役割を表す語)をラベルとして付与する手法を提案する．手法の概要を図 1 に示す．

2.1 職業関連情報候補抽出

職業関連情報候補抽出処理では，Web ページの中から HTML 構造に着目して人物に関連する名詞を抽出，その名詞からヒューリスティックを用いて職業関連情報の候補を抽出する．

2.1.1 名詞抽出

まず，HTML 構造に着目し，以下の 5 つの要素内に含まれる部分から，形態素解析を用いて名詞（単名詞または複合名詞）を抽出する．

1. 氏名が出現する p 要素（段落)内
2. 氏名が出現する tr 要素（表の列)内
3. 氏名が 1 行目または 2 行目に出現する table 要素(表）内
4. 氏名が出現する title 要素内
5. 氏名が出現する h1～h3 要素（見出し）内

次に，氏名の位置関係に着目し，以下の名詞を抽出する．

6. 氏名の直後に出現する丸括弧内の名詞
7. 氏名の直前と直後に出現する名詞

最後に，次節における著作候補の抽出のため，正規表現により ISBN 番号を抽出し，名詞に加えておく．

2.1.2 職業関連情報候補判定

得られた名詞について，ヒューリスティックを用いて職業関連情報の候補かどうかを判定する．判定は，職業を表す語，著作と役割を表す語，所属と役職を表す語の順番で行う．

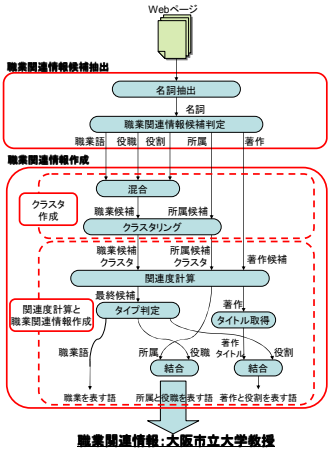

図 1 提案手法の概要

(1) 職業を表す語

名詞の語尾に着目した 17 種類のヒューリスティックを選定した．いずれかに合致する名詞を，職業を表す語（以下，職業語）の候補（以下，職業語候補）とする．選定したヒューリスティックを以下に示す．

- 「士」で終わる名詞
- カタカナ 4 文字以上で構成され，かつ語尾が「ー」で終わる名詞

(2) 所属と役職を表す語

所属と役職の 2 つに分けて判定する．

所属を表わす語（以下，所属語）の候補（以下，所属候補）の判定方法は 2 種類考案した．

まず，固有表現抽出システム NExT[1]に付属する組織名辞書を利用して判定する．以下に判定方法を示す．

- 「NExT の組織名辞書に含まれる 2 文字以上の語」が語尾につき，「合致した組織名辞書の語」の文字数+2 以上の文字数の名詞

次に 14 種類のヒューリスティックを選定した．以下に例を示す．

- 「(株)」で始まり，かつ 5 文字以上で構成される名詞
- 「株式会社」で始まり，かつ 6 文字以上で構成される名詞

役職に関しては，(1)と同じように，名詞の語尾に着目し，15 種類のヒューリスティックを選定し，合致するものを役職候補とする．以下に例を示す．

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

[1] http://www.ai.info.mie-u.ac.jp/~next/next.html

- 「員」で終わる名詞
- 「教授」で終わる名詞

(3) 著作と役割を表す語

著作と役割の 2 つに分けて判定する.

著作については, 前節で抽出した ISBN 番号をそのまま著作候補とする.

役割については, 以下の 3 つのヒューリスティックを用いて判定し, 役割候補とする.

- 「著者」で終わる名詞
- 「編者」で終わる名詞
- 「訳者」で終わる名詞

2.2 職業関連情報作成

職業関連情報の候補から職業関連情報を作成する. 著作候補を除いてクラスタ作成を行った後, 職業関連情報を作成する.

2.2.1 クラスタ作成

表記のゆれを吸収するために, 語尾に着目したクラスタリングを行い, 著作候補を除いて同義のクラスタを作成する.

職業語候補, 所属候補, 役職候補の 3 つを混ぜて職業候補クラスタを作成し, 所属候補から所属クラスタを作成する.

職業候補のクラスタリング手法を図 2 に示す.

Step.1 候補の出現頻度を計算する. 得られた候補と計算の結果得られた出現頻度の組を持つリスト(以下, 候補リスト)を作成する.
Step.2 候補リストのうち, 最も構成文字数が少ないものを選択し, その候補と出現頻度の組のみを含むクラスタを作成する.
Step.3 クラスタに属さない全て候補の語尾と, **Step.2**で選択した候補をパターンマッチにより比較する. **Step.2**で選択した候補が含まれれば**Step.2**で作成したクラスタに追加する.
Step.4 クラスタに追加した候補を候補リストから削除する.
Step.5 候補リストの中の候補がなくなるまで, **Step.2**から**Step.4**を実行する.

図 2 職業候補クラスタ作成手法

所属候補のクラスタリング手法は, 職業候補クラスタ作成手法(図 2)の Step.3 のみが異なる(図 3).

Step.3 クラスタに属さない全て候補の語頭と, **Step.2**で選択した候補をパターンマッチにより比較する. **Step.2**で選択した候補が含まれれば**Step.2**で作成したクラスタに追加する.

図 3 所属候補クラスタ作成手法の Step.3

2.2.2 関連度計算と職業関連情報作成

職業候補クラスタから, クラスタ内に含まれる候補の出現頻度の合計が最も多いクラスタを選択する. 最頻度が複数存在する場合は, 複数選択する. 選択されたクラスタの中に含まれる候補が 1 つのみの場合, その候補を最終候補とする.

候補が複数だった場合(複数候補と呼ぶ)に 1 つを選択する必要がある. 提案手法では, データ外部の情報に着目し, [森 05]の検索エンジンを用いたスコア計算法を用いる. 我々は, 森らの手法を人物と複数候補との関連度計算に応用し, 関連度の最も高い候補を最終候補とする.

氏名 n と複数候補 v の関連度 $J(n,v)$を以下のように求める.

$$J(n,v)=\frac{|N\cap V|}{|N|+|V|-|N\cap V|}$$

なお, $|N|$は検索クエリを n として Web 検索エンジンから得られるヒット数, $|V|$は検索クエリを v として得られるヒット数, $|N\cap V|$は n と v の AND 検索にて得られるヒット数である.

最終候補には, 職業語, 役職, 役割の 3 種類がある. 最終候補のタイプを判定し, タイプ毎に異なる処理を行い, 職業関連情報を作成する.

最終候補が職業語である場合には, それを職業関連情報とする.

最終候補が役職であれば所属と結合する. しかし,「大阪市立大学教授」などのように, 役職に所属が含まれている場合もある. そのため, 役職に所属が含まれるかどうかを判定し, 含まれていなければ所属を結合する. 判定には 2.1.2 節の所属の判定で用いている語を使用する. いずれかの語が最終候補に含まれていれば, 所属が既に含まれていると判定し, 最終候補を職業関連情報とする.

所属が含まれていないと判定された場合, 所属候補クラスタから所属を得る. まず, クラスタ内の候補の出現頻度の合計が最も多いクラスタを選択する. 選択されたクラスタ内に所属候補が 1 つであれば, その所属候補を最終候補と結合し, 職業関連情報とする. クラスタ内に所属候補が複数ある場合や最頻度のクラスタが複数ある場合は関連度を計算する. 関連度の最も高い所属候補を最終候補と結合し, 職業関連情報とする. 関連度 $J(n,o,v)$を,

$$J(n,o,v)=\frac{|N\cap O\cap V|}{|N\cap V|+|O|-|N\cap O\cap V|}$$

と定義する. n は氏名, o は所属候補, v は最終候補(役職)である. $|N\cap V|$は n と v の AND 検索にて得られるヒット数, $|O|$ は o で得られるヒット数, $|N\cap O\cap V|$は, n と o と v の AND 検索にて得られるヒット数である.

最終候補が役割であれば, 著作候補から関連の高い著作を選択して結合する. 著作候補の関連度の計算には, 著作候補(ISBN 番号)と氏名との間の文字数(以下, 文字距離)を用いて, 文字距離が最も短い ISBN 番号を 1 つ選択する. 選択された ISBN 番号を用いて著作タイトルを取得する. 著作タイトルと最終候補(役割)を結合して職業関連情報を作成する.

3. 関連研究

本研究の目的は Web 上の同姓同名人物の識別を容易にすることである. 代表的な先行研究として, [Wan 05]がある. [Wan 05]では人物のラベルとして, 肩書を用いている. 本研究では, 職業関連情報をラベルとしている.

Web 上の同姓同名人物に関する研究は, 同姓同名人物の分離([佐藤 05]など)がほとんどである. 本研究では, 同姓同名人物の分離を対象としていない.

4. おわりに

本研究では, 同姓同名人物の識別を容易にするために, 職業関連情報を抽出する手法を提案した. 本研究における職業関連情報とは, 職業を表す語, 所属と役職を表す語, 著作と役割を表す語, である. 職業関連情報候補抽出処理で, 職業関連情報の候補を抽出し, 職業関連情報作成処理では得られた候補から職業関連情報を作成する.

今後は, 本手法で抽出した職業関連情報の有用性を確認するために実験を行う予定である.

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, 3G8-1(2007)

[森 05] 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 (2005)

[Wan 05] Wan, X., Gao, J., Li, M., and Ding, B.: Person Resolution in Person Search Results: WebHawk, in Proc. of.CIKM'05 (2005)

[佐藤 05] 佐藤進也, 風間一洋, 福田健介, 村上健一郎:実世界指向 Web マイニングによる同姓同名人物の分離, 情報処理学会論文誌:データベース, Vol.46 No. SIG 8 (2005)