3G8-1

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

A system that distinguishes different people with identical names on the Web and displays a list by attribute information

上田 洋[*1]
Hiroshi UEDA

村上 晴美[*2]
Harumi MURAKAMI

[*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

Distinguishing different people with identical names is crucial when searching by name. Much research has classified different people on the Web by concentrating on classification methods. In this paper, we focus on a user interface that displays different people who have been distinguished to simplify user selection. We developed a prototype that distinguishes different people with identical names on the Web. It displays a list of people by related prefectures, vocations, and keywords.

1. はじめに

人名を用いた情報検索においては, 検索結果を同姓同名人物毎に分類することが課題である. そこで, Web 検索においては, 人名による検索結果を同姓同名人物に自動分離する研究(たとえば[木村 06])が行われている. また, Web マイニングにおいては, オブジェクト識別問題の一つとして, Web 上の同姓同名人物分離の研究が行われている(たとえば[佐藤 05]). これらの研究の多くは, クラスタリング, Web の構造情報の利用などをはじめとして, 手法に焦点があてられている.

本研究は, Web 検索において, 分離された同姓同名人物の選択を容易にするためのユーザインタフェースを実現することを目的とする. 該当人物に関連する属性情報として, 地方, 職業, キーワードを表示するシステムを提案する.

以下, 2 節でシステムの概要, 3 節で実行例, 4 節で関連研究について述べる.

2. システム

システムは, 氏名を用いた Web 検索結果を同姓同名人物に分離し, (a) 該当人物に関連する地方, 職業, キーワードを表形式で表示する人物属性情報一覧画面, (b) 該当人物のクラスタを 2 次元空間にマッピングする人物類似度空間画面, を表示する.

以下では, (a), (b)共通処理として, 同姓同名人物の分離, 関連地方・職業・キーワードの抽出, (b)のみの処理として, 同姓同名人物の可視化, について述べる.

2.1 同姓同名人物の分離

氏名を入力文字列として, Google Web APIs を用いて Web ページを取得する. 本システムでは, 非階層型クラスタリングの一種である単一パス法に改良を加えたもの(以下, 改良単一パス法)を用いて同姓同名人物の分離を行う.

改良単一パス法の手順は以下の通りである.

1. 閾値を設定する.
2. 最初の要素を選択し, 最初のクラスタとする.
3. 一番新しく作成したクラスタと, クラスタに含まれる以外の要素との類似度を全て計算する.
4. 設定閾値を超えたものの中で最も高い類似度の要素をクラスタに入れる. もし, 設定閾値を超えるものがなければ, まだクラスタに属していない要素を選択し, その要素を含む新たなクラスタを作成する.
5. クラスタに属さない要素がなくなるまで, 3.と 4.を繰り返す.
6. 5. までの処理で得られたクラスタのうち, 要素の数が1のクラスタを削除する.
7. 残ったクラスタに対して, 各クラスタ間の類似度を計算し, 設定閾値を超えるクラスタ同士を結合する.

重み付けは, TF・IDF 法にて行い, 類似度には, ベクトルの余弦を用いて計算する.

2.2 関連地方の提示

本システムでは, 該当人物に関連する都道府県を関連地方として 1 つ提示する. 関連地方は 2.1 節で該当人物のクラスタに含まれる Web ページから抽出する. 入力氏名の前後(入力氏名の前部 20 文字, 後部 100 文字)の文字列の中に含まれる都道府県名を抽出し, 最頻度の都道府県名を提示する. 最頻度の都道府県名が複数存在した場合, 入力氏名の最も近くに存在するものを提示する.

2.3 関連職業の提示

該当人物に関連する職業を関連職業として1つ提示する.

事前処理として, 関連職業の抽出に用いる職業辞書を作成する. 職業辞書は,「職業名, スコア順(後述)に並べられた 20 語」で構成される. 職業辞書の作成方法は以下の通りである.

Wikipedia の職業一覧のページ[1]より, 職業名を抽出する. 得られた職業名にリンクされているページを取得する. 2007 年 4 月 10 日現在, 203 の職業がリンクされている. 各ページの概要に当たる部分を抽出する. 得られた文字列について形態素解析を行う. その結果, 名詞と判定され, 2 文字以上から構成される語を抽出する. 得られた語について TF・IDF 法で重み付けし, スコアの高い 20 語を各職業の関連する語として職業辞書に登録する.

関連職業の抽出の手順は以下の通りである.

まず, 203 の職業名が入力氏名の近くの文字列に含まれるかどうか調査する. 1 件だけ含まれる場合, その職業名を関連職業として提示する. 複数ある場合は, 最も入力氏名に近い職業

連絡先: 上田 洋, 大阪市立大学, 06-6605-3375 (村上研究室), d06tb001@ex.media.osaka-cu.ac.jp

[1] http://ja.wikipedia.org/wiki/%E8%81%B7%E6%A5%AD%E4%B8%80%E8%A6%A7

を提示する. もし, 職業名が存在しなければ, 各職業の関連語を用いて, 関連職業抽出を行う.

次に, 入力氏名の近くの文字列に含まれる職業辞書の語を抽出し, 出現頻度を計算する. 各職業の 20 語について頻度にランク毎に重みをつけ, 各語のスコアを計算する. 20 語のスコアの合計を各職業のスコアとする. スコアの最も高い職業を 1 件提示する. なお, 最高スコアの職業が複数存在した場合, 各語と入力氏名との間の語数の平均が最も小さいものを提示する.

2.4 関連キーワードの提示

関連キーワードは, 該当人物に関連のあるキーワードである. 本システムでは, 10 語提示する.

2.2 節, 2.3 節と同じく入力氏名の近くの文字列を用いて関連キーワードを抽出する. 抽出には, [中川 03]の手法を用いる. [中川 03]では, 単名詞の出現頻度と連接頻度を用いた専門用語抽出の手法を提案している. 本システムでは, 関連キーワードとして複合名詞を提示する. 複合名詞のスコアを計算し, 高スコアの 10 語を提示する.

複合名詞のスコアを

$$FLR(CN) = f(CN) \cdot LR(CN)$$

と定義した. CN は単名詞 $N_1,N_2,\ldots N_L$ の順に連接する複合名詞, $f(CN)$ は CN の文書頻度(DF)である.

$LR(CN)$ は

$$LR(CN) = (\prod_{i=1}^{L}(FL(N_i)+1)(FR(N_i)+1))^{1/2L}$$

とする. $FL(N_i)$ は単名詞 N_i の左方の連接する名詞の異なり数, $FR(N_i)$ は右方の連接名詞の異なり数である.

2.5 同姓同名人物の可視化

本システムでは, 各クラスタを各同姓同名人物とみなし, 2 次元空間にマッピングする.

まず, 全ての Web ページに対してベクトルデータを作成する. 作成したベクトルデータ群に対し, 潜在的意味インデキシングを用いて次元数を 2 に圧縮する. 圧縮されたデータを用いて, 各 Web ページの座標を算出する. クラスタ内の Web ページの座標を元に, クラスタの座標を算出し, 各人物のマッピングを行う.

3. 実行例

図 1, 2 に, 入力氏名「村上 晴美」により表示される人物属性情報一覧画面と人物類似度空間画面の例を示す.

たとえば, 1 人目のクラスタの属性情報として, 地方が「大阪府」, 職業が「教授」, キーワードとして「情報検索」「学術情報総合センター」「情報検索システム」などが表示されているが, これは第二著者のクラスタである.

4. 関連研究

Web ページから人物属性情報を抽出する研究には, [森 05]や[山本 00]の研究などがある. [森 05]では, 検索エンジンを利用して得られた語の共起情報を統計的に処理し, 研究者のキーワードを抽出する手法を提案している. [山本 00]では, 検索エンジンとハイパーリンクを用いて, 与えられた人名から人物紹介のテキスト情報を抽出する手法を提案している.

本研究では, 関連地方・職業・キーワードを人物属性情報として抽出した. 関連地方・職業に関しては, あらかじめ作成した辞書ファイルを用いて抽出を行った. 関連キーワードについては, Web ページに出現する複合名詞を用いて, 高スコアの複合名詞を抽出した.

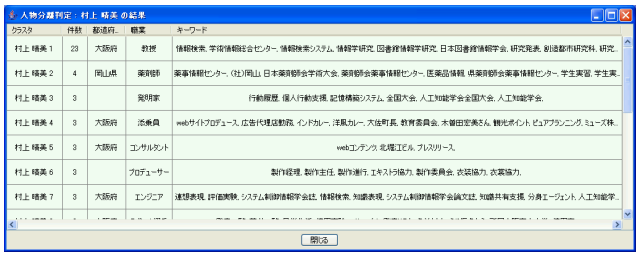

図 1: 人物属性情報一覧画面

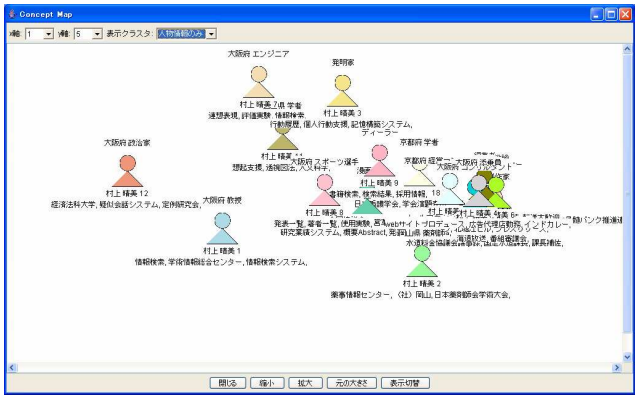

図 2: 人物類似度空間画面

5. おわりに

本研究では, 人名による Web 検索結果を同姓同名人物に分離し, 分離された人物毎に関連する属性情報(地方, 職業, キーワード)を表示するシステムを試作した.

システムは, (a) 該当人物に関連する地方, 職業, キーワードを表形式で表示する人物属性情報一覧画面, (b) 該当人物のクラスタを 2 次元空間にマッピングする人物類似度空間画面, を表示する.

本システムは, Web 検索結果の分類インタフェースとしてだけではなく, 該当人物の特徴の理解にも役立つのではないと考えている.

今後の課題として, 同姓同名分離や属性情報提示処理の精度向上や, ユーザ実験などがあげられる.

参考文献

[木村 06] 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, DEWS2006, 2C-i11, 2006.

[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.

[中川 03] 中川裕志, 森辰則, 湯本紘彰: 出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理, Vol.10 No.1, 2003.

[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.

[山本 00] 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, ICS-119-24, 2000.