Web 人名検索の要約と可視化を目指して

Toward Summarization and Visualization of Web People Search Results

村上 晴美[*1]
Harumi MURAKAMI

上田 洋[*2]
Hiroshi UEDA

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

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

We conduct research to understand people using texts and databases. In this research, we present a project that summarizes and visualizes Web people search results and report the progress of two of its aspects. (1) We investigated algorithms for distinguishing individuals with identical names and extracted attribute information of individuals, including keywords, prefectures, addresses, vocations, and vocation-related information. (2) We developed prototypes to display separated individuals by lists, maps, and two-dimensional space interfaces.

1. はじめに

筆者らの研究室では，テキストやデータからの人物の理解に関する研究を行っており，その一つとして，Web 人名検索結果の要約と可視化研究を行っている．本稿では，Web 人名検索結果の要約と可視化プロジェクトの概要と進捗について述べる．

2. プロジェクト概要

近年，Web 検索の約 3 割は人名検索と言われとおり，人名検索においては人名の曖昧性解消（日本では同姓同名人物の分離）が重要な課題となってきている．

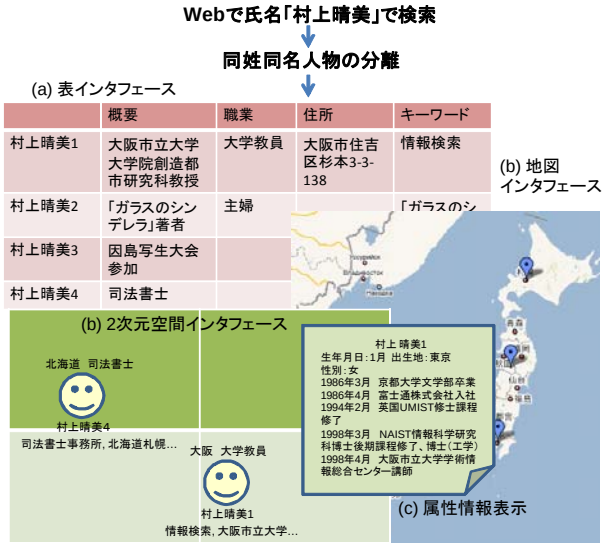

図 1：プロジェクト概要

本プロジェクトの目的は，Web 上の人名検索においてユーザによる人物の理解と選択を支援するシステムの開発である．

人名検索結果の Web ページを同姓同名人物毎に分類して人物クラスタを作成する．ユーザによる人物の理解と選択を助けるために，人物クラスタ毎に，(a) 概要，キーワード，職業，住所などの属性情報を提示する表インタフェースと，(b) 2 次元空間上に人物アイコンを表示する 2 時限空間インタフェースと，(b) 人物クラスタの代表的な位置情報を取得して地図上に人物アイコンを表示する地図インタフェースを開発する．(c) ユーザが人物クラスタを選択すると，人物に応じた属性情報を取捨選択して表示する．図 1 にプロジェクト概要を示す．

本研究の特徴は，(1) 人物の理解が目的であること，(2) 人物に関連するすべての属性情報を抽出するのではなくそれらを統合して最適な属性情報を提示すること，(3) インタフェースに焦点をあてていることである．本稿では「抽出された属性情報を統合して，人物の理解や選択に有用な情報を提示すること」を「要約」と呼ぶ．要約は抽出ではなく，抽出された情報や他の情報源から取得した情報の統合や付与などを含む．

以下，3-7 節では，これまでの進捗を報告する．

3. 初期プロトタイプの試作と表インタフェースの実験

2006-2007 年に，Google から人名検索結果を取得し，結果として得た Web ページを同姓同名人物に分離し，(a) 人物属性情報（職業，都道府県名，キーワード）を表形式で表示するインタフェースと，(b) 人物アイコンを 2 次元空間上に表示するインタフェースを試作した[上田 07]．初期プロトタイプを図 2 に示す．

予備的な実験の結果，同姓同名人物の分離手法の精度に問題があること（4 節にて述べる），表インタフェースにおける人物選択においてキーワードと職業が有用であることが示された．

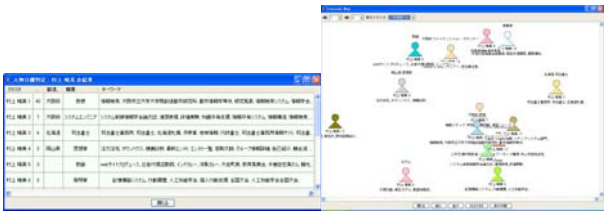

図 2：初期プロトタイプ

4. 同姓同名人物の分離

同姓同名人物の分離においては，当初（2006 年），特徴語の階層型クラスタリング手法が主流であると考えられたが，我々は「人間が行う分類」に近いと思われる方法を検討するため，特徴語の非階層型クラスタリング（単一パス法）に基づく分離を試行した．すなわち，検索結果の最上位からチェックして，類似するものを一つのクラスタに，類似しないものを別のクラスタに分ける

連絡先：村上 晴美，大阪市立大学大学院創造都市研究科，
〒558-8585 大阪市住吉区杉本 3-3-138，06-6605-3375，
harumi@media.osaka-cu.ac.jp

という方法である. 結果は細かく分けすぎになってしまうという問題点があった[上田 07]. この方法の改善策として, 類似度の判定方法の改善や, 分離されたクラスタを統合していくアプローチが考えられる.

別のアプローチとして, [片岡 08]では, クエリでない人名の共起に基づく中間クラスタを作成後に, 特徴語を用いた階層型クラスタリングを行うという方法を検討した. 本研究は Wan ら[Wan95]の改良研究と位置付けられる. 性能評価の結果, Wan らとほぼ同等であり, WePS[Artiles 07]との数値比較を行ったところ, 2 位に位置付けられた[片岡 09].

5. 職業関連情報の付与

[上田 07]においては, 日本には適当な職業辞書がないという問題があり, Wikipedia の職業ページから職業辞書を作成し, 人物クラスタに付与を試みた. また, たとえば, 第一著者の職業として,「研究者」「教員」「大学教授」を付与すればよいのか, という問題や,「研究者」よりも「大阪市立大学大学院創造都市研究科教授」の方が有用なのでは, という疑問が残った.

[上田 08]では, 人物クラスタに最適な「職業関連情報」を一つ付与する手法を提案した. これは厳密に職業と定義される語だけではなく, 幅広く職業と考えられる語や, 職業の推定に有用と思われる語も含めた, 職業に関連する情報である. 提案手法は, (a) HTML 構造と簡単なヒューリスティックを用いた職業関連情報候補抽出, (b) 出現頻度, 同義クラスタ作成, Web 検索エンジンを用いた職業関連情報作成から構成される. 手法の特徴は, 厳密に定義された職業を出現頻度に応じて付与するのではなく, 人物の識別に有用と思われる職業関連情報を文脈に応じて生成して付与する点である. 特別な職業辞書が不要であることも利点である. たとえば, 漫画「美味しんぼ」の主人公山岡士郎の人物クラスタに対して「東西新聞文化部記者」と付与される[Ueda 09, 上田 09a].

6. 位置情報の付与と地図インタフェースの試作

人物選択のための有用な属性情報として地名(住所)が存在する. [高守 09a] においては, 表インタフェースにおける住所の取得と, 地図インタフェースの作成を行うために, 人物クラスタに最適な位置情報を一つ付与した.

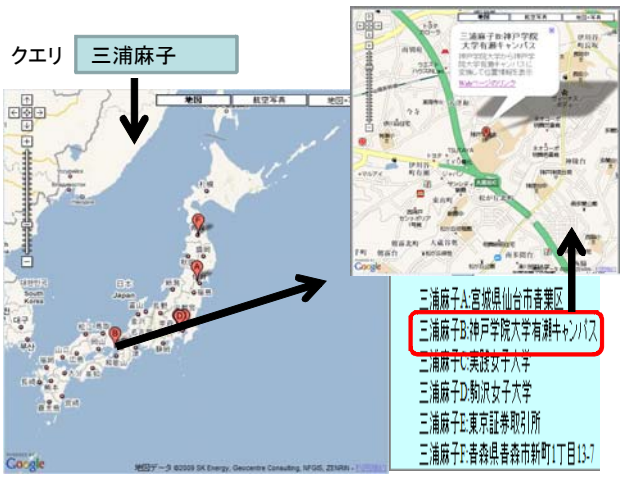

図 3：位置情報の付与と地図インタフェース

最適な位置情報として, 該当人物の現在の勤務先や自宅の付与を目指した. 提案手法の特徴は, Yahoo!ローカル検索の利用, ランドマークへの着目, Web 検索エンジンランキングと文字距離を用いた位置情報候補の取得である. たとえば, 神戸学院大学の三浦麻子氏の人物クラスタに対して前職の「大阪大学」ではなく現職の「神戸学院大学有瀬キャンパス」の位置情報が付与される[1]. また, 提案手法を用いた地図インタフェースのプロトタイプ(図 3)を試作した[高守 09b].

7. イベント情報の抽出

人物クラスタを選択したときに表示される人物属性情報には多様な方法が考えられる. 最も一般的なものがリスト形式である. [上田 09b]では, 利用者にとってなじみがある履歴書を人物クラスタから作成することを目指し, 履歴書に記載されているイベント情報(年月日と出来事を含む文)を抽出している.

8. おわりに

Web 人名検索結果の要約と可視化を目指す研究についてプロジェクト概要を述べ, これまでの進捗を報告した. 現在は, キーワード抽出手法と履歴書作成手法を検討中である.

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, 2007 年度人工知能学会全国大会(第 21 回)論文集, 2007.

[片岡 08] 片岡 真一, 上田 洋, 村上 晴美, 辰己 昭治: 人物名に着目した二段階クラスタリングによる同姓同名人物の分離, 2008 年度人工知能学会全国大会(第 22 回)論文集, 2008.

[Wan 95] X. Wan, J. Gao, M. Li, and B. Ding, "Person Resolution in Person Search Results: WebHawk", CIKM2005, Proceedings of the Fourteenth ACM Conference on Information and Knowledge Management, 2005, pp. 163-170.

[Artiles 07] J. Artiles, J. Gonzalo, and S. Sekine: The SemEval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task, Proc. SemEval 2007, ACL, 2007.

[片岡 09] 片岡 真一, 人物名に着目した二段階クラスタリングによる同姓同名人物の分離, 大阪市立大学大学院創造都市研究科都市情報学専攻修士論文, 2009.

[上田 08] 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, 2008 年度人工知能学会全国大会(第 22 回)論文集, 2008.

[Ueda 09] Hiroshi Ueda, Harumi Murakami, Shoji Tatsumi: Assigning Vocation-Related Information for Person Clusters from Web People Search Results, Proceedings GCIS 2009, 2009 (to appear).

[上田 09a] 上田 洋, 村上 晴美: 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 09a] 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 回全国大会(平成 21 年)講演論文集, 1, 625-626, 2009.

[高守 09b] 高守 雄也, Web ページからの人物に関する位置情報の抽出, 大阪市立大学大学院創造都市研究科都市情報学専攻修士論文, 2009.

[上田 09b] 上田 洋, 村上 晴美, 辰巳 昭治, Web からの履歴書作成のためのイベント情報の抽出, 2009 年度人工知能学会全国大会(第 23 回)論文集, 2009(予定).

[1] 2009 年 4 月に転出のため, 現在の所属は関西学院大学である.