LD-005

文書内の人名の個人特定に関する研究 —「山田さん問題」の解決手法とその評価—

Research on Person Identification for a Person Name in a Document --- A Method of Solving "Yamada-san Problem" and Its Evaluation ---

松平　正樹　Masaki Matsudaira
上田　俊夫　Toshio Ueda
大沼　宏行　Hiroyuki Ohnuma
渕上　正睦　Masachika Fuchigami
森田　幸伯　Yukihiro Morita

1. はじめに

情報を収集する際，必要な情報が不必要な情報に埋もれてしまう情報洪水の問題に対して，我々は，文書からキーワードを抽出し，その意味属性に応じて情報を収集，意味的に整理して出力するシステムを開発している[松平 03].

システムは，以下のような利用シーンを想定し，文書閲覧時に，その文脈で情報を収集することが目的である.

i. 会議報告書から他社製品名を抽出し，その製品に関する情報を収集する
ii. メールから人名を抽出し，電話／メールで連絡をとる
iii. 会議開催案内メールから日付・時間を抽出し，スケジュールを確認する

しかしながら，文書内のキーワードと情報との対応づけにおいて，特に姓だけで出現する人名の個人特定の問題が生じる（「山田さん問題」と呼ぶ）[松平 04]. 我々の調査では，イントラネット文書の約 1/3 の人名が姓だけで出現しているという結果を得ており，前後に組織名を伴っていても異動や組織変更により変更されているケースもあり，この問題の影響は大きいと考えている.

本報告では，まず 2 章で「山田さん問題」について概説し，3 章で我々の提案する解決手法について説明する．4 章では，イントラネット文書を対象におこなった，姓から個人を特定する評価実験の内容および評価結果について述べる.

2. 山田さん問題

個人を特定するためには，どのような情報が必要であろうか？

Semantic Web の世界では，データに対して RDF（Resource Description Framework）によるメタデータを付与し，リソースの URI によって個人を識別することができる．例えば，SHOE（Simple HTML Ontology Extensions）プロジェクトでは，HTML の拡張としてメタデータを記述するためのタグを規定しており[Heflin 98, SHOE]，FOAF（Friend of a Friend）プロジェクトでは，ある人の友人を，メールボックスを利用して特定している[FOAF]．また，知識ベースに RDF 形式のデータを保持し，アーティスト等の人物に関連する情報を複数のサイトから収集する TAP Semantic Search の研究もある[Guha 03, TAP]．企業内の従業員の場合は，氏名や従業員番号，メールアドレス等で特定することができるだろう．メタタグを付与するためのツールもいくつか開発されている.

しかしながら，一般的には大量にある既存の文書を対象にしたいという要求が強く，しかも，前述したように姓だけで出現する場合も多く，文書内にメールアドレスや従業員番号を伴うとは限らない．その場合，単に「山田」で社内の従業員データベースを検索すると 100 件以上の結果が見つかり，何らかの制約が必要である．本稿では，

- 限定された範囲での個人情報データベースがあることを想定し，
- その範囲内で姓から個人名の候補を絞る

という問題を「山田さん問題」と定義する．図 1 に例を示す.

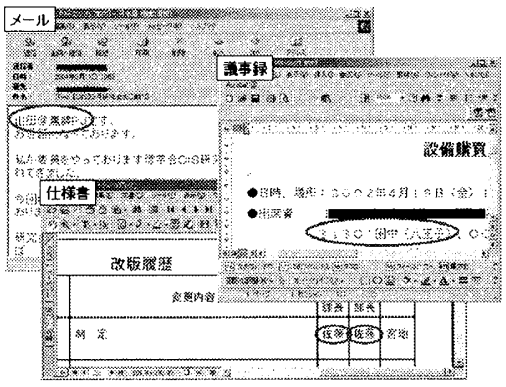

図 1　「山田さん問題」の例

沖電気工業（株）　研究開発本部
Oki Electric Industry, Co., Ltd. Research and Development Division

2.1. 出現パターン分析

「山田さん問題」に対して，イントラネット文書の出現パターンを分析した結果，議事録，仕様書，管理表，組織ホームページ，メール等の文書カテゴリごとに出現パターンに特徴があることが判明した[松平 04]．以下に例を示す．

議事録

出席者および発言者として記述される．組織名あるいは役職名を伴うものが多い．

仕様書

改版履歴の記入者，承認者として記述される．組織名を伴って記述されるか，伴わない場合は仕様書の表紙ページに組織名がある．

管理表

1行内に組織名，担当製品，担当顧客等の情報が対応づけられている．電話番号，メールアドレスが記述されている場合もある．

3. 問題の解決方式

3.1. 従来研究

「山田さん問題」に関連する従来研究として，文書内から人名を抽出し，職業や出身地等の人物情報と対応づけるという研究がある[西野 98]．西野らは，職業「作詞家」に対して新聞記事から人名を抽出する実験をおこなっており，再現率98%，適合率73%と良好な結果が報告されている．しかしながら，姓ではなく姓名を対象としており，また，同姓同名を区別するまでには至っていない．

一方，オントロジーを機械翻訳の多義性解消に応用した研究がある[Kang 01]．この研究では，シソーラス辞書から概念間を関係づけたオントロジーを構築し，1文内に出現する単語の概念を制約として利用することにより，高い精度で多義性を解消できるという成果を得ている．しかしながら，「山田さん問題」はひとつの概念（すなわち，「人」という概念）内で実体を特定するものであり，複数の概念の候補からひとつを特定する多義性解消とは，問題の次元が異なると考えられる．

3.2. 提案方式

我々は，出現パターン分析に基づく文書知識と，人に関するオントロジーを利用して，文書内のキーワードを人の属性の制約と見なし，制約の種類，制約と人名の出現位置，文書の種類に応じて制約の強さを調節し，さらに推論をおこなって候補を絞り込むアプローチをとる．全体の流れを図2に示す．

固有表現抽出では，人に関するオントロジーの属性項目になり得るキーワードを抽出する．抽出する属性の一部を以下に示す．

Person_Lname	姓
Person_Fname	名
Organization_Name	組織名

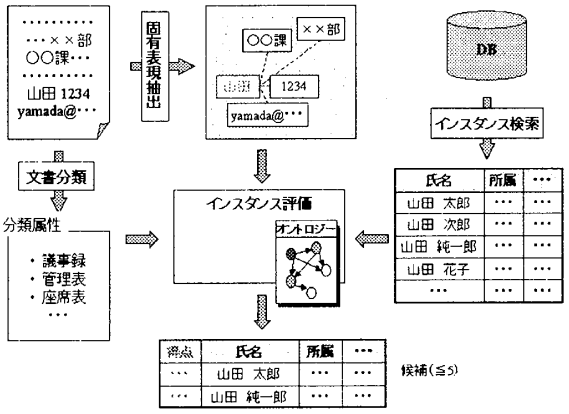

図2 全体の流れ

SubOrganization_Name	サブ組織名
Occupation	役職
Telephone_Number	電話番号
Mail_Address	メールアドレス
Product_Name	製品名
Technology_Name	技術名
・・・	

文書分類では，URLや文書タイトル，ファイル形式，本文内のキーワードのヒューリスティックルールにより，元の文書を議事録，仕様書，管理表，メール等の文書カテゴリに分類する．SVM等の学習法による文書分類と異なり，後述する制約の重みを補正するために，分野には依存しない文書の形式によって分類している．

インスタンス検索は，姓から従業員データベースや顧客データベースを検索し，結果をインスタンス候補として出力する．

オントロジーおよびインスタンス評価については，次節で詳述する．

3.2.1. オントロジー

「山田さん問題」を解決するためのオントロジーは，人に関連する概念と概念の関係を定義したもので，我々が独自に構築した．オントロジーの一部を図3に示す．

図3上部は，Person（人）クラスとPerson_LName（姓）等の属性項目、Personクラスと Employee（従業員）クラスの上位・下位関係，および属性項目とSubOrganization（サブ組織）等のベースクラスの関係を示している．ベースクラスは，属性項目に入る値の型およびクラスである．

図3下部はEmployee（従業員）クラスのインスタンスの例であり，「山田」「太郎」「営業部」「03-1234-5678」という属性を持つ従業員クラスの1つのインスタンスを示している．

3.2.2. インスタンス評価

インスタンス評価では，インスタンス検索により得られた各インスタンス候補に対して，文書内のキ

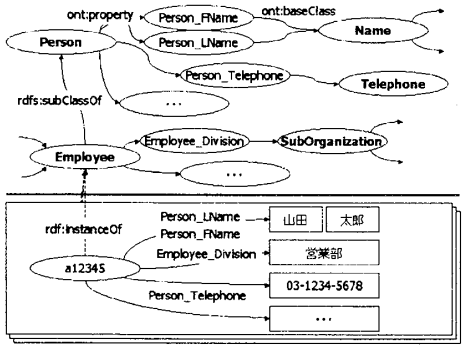

図３　オントロジー

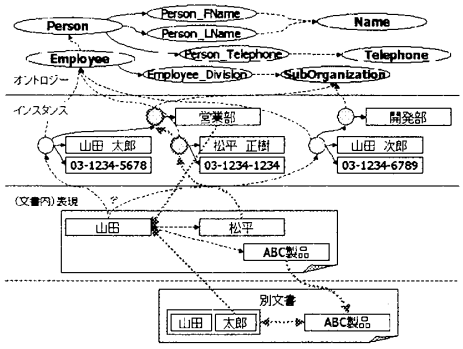

図４　制約の１次推論

ーワードを制約と見なし，その属性，位置情報，文書カテゴリを用いて各インスタンス候補の評価をおこなう．インスタンス評価のアルゴリズムを以下に示す．

キーワードを直接制約として，あるいは，キーワードから１次推論によって制約を導出し，マッチする各制約に応じた得点を加算する．その後，各インスタンス候補の得点に応じて絞込みの結果を出力する．一定の個数 N に絞れなかった場合は，候補を絞

```
姓にマッチするインスタンスを検索
for 各インスタンス候補について
  for 各キーワードについて
    if インスタンス候補と制約がマッチ
      インスタンス候補に得点を加算
    endif
if 最高得点から一定値 Sdist 以内の候補の数がN以内
  インスタンス候補を出力して終了
else
  候補絞れずとして終了
endif
```

れなかった旨出力する．

制約の１次推論

１回だけオントロジー上の推論をおこない，制約を導出するものである．図４に１次推論の概念を示す．１次推論は，定義したオントロジー内でおこなう推論と，別文書からインスタンスの関係を抽出しておこなう推論がある．例えば，前者は議事録では組織名を伴わない人名は同じ組織に属している可能性が高いという推測から所属を推論するものであり，後者は製品名や技術名、顧客名等から別文書に出現した本人の氏名、電話番号等を推論するものである．

得点計算

得点は次式によって与える．

$$S = \Sigma S_B(C(K_i), C_{PS})*W(C_D,pos(K_i),pos(K_{ps})) + \Sigma (S_B(C(X),C_{PS})*W(C_D,pos(X),pos(K_{ps})) * S_B(C(K_i),C(X))*W(C_X,pos(K_i),pos(X)))$$

得点 S は，各キーワード K_i に対して，直接制約として，制約の属性 $C(K_i)$ に応じた基礎点 S_B と，文書カテゴリ C_D，キーワードの出現位置 $pos(K_i)$，姓 K_{ps} の出現位置 $pos(K_{ps})$から求められる重みづけ W の積（第１項），および１次推論による制約として，推論された制約の属性 C(X)に応じた基礎点および重みづけ，推論のため媒体キーワードとなる X をキーワード K_i の推論とみなした場合の基礎点，重みづけの積（第２項）の和によって求める．

基礎点 S_B は（$0 \leqq S_B \leqq 1$），個人を特定できる可能性を数値化したもので，属性値の一意性や変更頻度に依存する．重みづけ W（$0 \leqq W \leqq 1$）は，制約と姓の出現位置が近いほど高くするが，文書カテゴリに応じて補正している．これは，管理表のように同一行内での制約しか有効ではない場合や，仕様書のように文書先頭の一定範囲内に出現する組織名を制約として利用すべき場合を考慮したものである．

4.　評価実験

4.1.　実験対象・条件

弊社のイントラネット文書および従業員データベースを対象に，日本人に多い「佐藤」「鈴木」「高橋」「田中」「山田」の５つの姓について文書検索をおこない，「山田さん問題」の生じる 173 文書の中から社外の人名や，座席表等，ほぼ内容が同等のものを対象外として，87 文書を選択した．

基礎点 S_B および重みづけ W は，統計的な手法で求めるべきであるが，今回の実験では，属性値の一意性（メールアドレスは個人で一意）や変更頻度（組織名は度々変更される）から経験的に設定した．

メールアドレス S_B=1.0
電話番号　　S_B=0.8
組織名　　S_B=0.5
役職　　S_B=0.3
管理表以外：前後 K 行の制約　W=±1-0.1*K
管理表：先頭から３行の組織名　W=0.5
……

また，出力する候補数の最大値 N は 5，最高得点との差 S_{dist} は 0.3 とした．MS Office ファイル，PDF ファイルは，テキストを抽出し，組織名の漢数字／算用数字，全角／半角等を正式表記に統一している．

比較のため，我々の提案方式（提案方式 2 とする）の他に，同一行および前後 1 行内のキーワードを直接制約としたベース方式 1，基礎点および重みづけを一定値とした（すなわち，マッチする制約の個数に依存する）直接制約のみのベース方式 2，ならびに基礎点および重みづけを考慮した推論なしの方式（提案方式 1 とする）について評価をおこなった．

4.2. 結果と考察

評価結果を表 1 および図 5 に示す．

候補正解とは，候補を 1 つに絞り，かつそれが正解である場合，候補内正解とは，複数（5 以内）の候補内に正解がある場合を示し，候補内に正解がない場合を不正解，候補を 5 以内に絞れなかった場合を候補絞れずとした．

表 1　評価結果

	Base1	Base2	提案 1	提案 2
正解	46.0%	54.0%	60.9%	73.6%
候補正解	40.2%	48.3%	54.0%	63.2%
候補内正解	5.7%	5.7%	6.9%	10.3%
不正解	4.6%	4.6%	3.4%	5.7%
候補絞れず	56.3%	41.4%	35.6%	20.7%

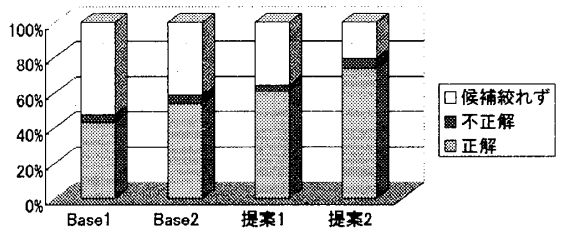

図 5　評価結果グラフ

結果から，ベース方式 1、ベース方式 2 と比較して，提案方式 1，さらに提案方式 2 は，候補を絞り込み，正解率が高いことがわかる．内容を分析すると，まず，ベース方式 1 では，離れた制約が作用せず，候補を挙げられない場合が多く，逆にベース方式 2 では，無関係の属性が制約として作用し，不必要な候補を取り上げるケースが目立つ．提案方式では，制約の属性に応じた基礎点を設け，文書カテゴリによる重みづけをおこなっていることにより，この問題が解消されている．

また，提案方式 2 では制約の 1 次推論により，例えば，制約がほとんど抽出できなかった議事録において，人名から組織名を制約として導出し，有効に作用している．

提案方式 2 で不正解，あるいは候補を絞れなかったものは，情報が旧く，抽出した組織名，製品名等が制約として作用しなかった場合が多い．文書の日付に応じた何らかの処理が必要である．

5. まとめと今後の課題

「山田さん問題」を定義し，オントロジーを利用して他のキーワードを属性の制約と見なし，制約の属性，文書カテゴリ，制約と人名の出現位置に応じて得点を計算することにより個人を特定する手法を提案した．提案手法では，制約が直接インスタンスの属性になっていない場合でも，1 次推論をおこなうことにより制約を導出している．イントラネットを対象とした評価実験は，従来の方式に比べて良好な結果が得られた．

今後は，経験的に設定した基礎点や重みを統計的な手法によって求めるとともに，社外の人名を含めた場合の問題解決や，同姓同名の個人特定問題（広義の「山田さん問題」）に取り組んでいく予定である．

参考文献

[FOAF] The Friend of a Friend (FOAF) Project: http://www.foaf-project.org/

[Guha 03] Guha, R., McCool, R., and Miller, E.: Semantic Search, The Twelfth International World Wide Web Conference (WWW2003), 2003

[Heflin 98] Heflin, J., Hendler, J., and Luke, S.: Reading Between the Lines: Using SHOE to Discover Implicit Knowledge from the Web. In AI and Information Integration. Papers from the 1998 Workshop. WS-98-14. AAAI Press, 1998

[Kang 01] Kang, S.J. and Lee, J.H.: Ontology-Based Word Sense Disambiguation by Using Semi-Automatically Constructed Ontology. 8th Machine Translation Summit (MT Summit Ⅷ). 2001

[SHOE] Simple HTML Ontology Extensions: http://www.cs.umd.edu/projects/plus/SHOE/

[TAP] TAP Semantic Search, http://tap.stanford.edu/

[西野 98] 西野，落谷: 新聞記事からの人物・企業情報の抽出，情報処理学会　自然言語処理研究報告 127-17, 1998

[松平 03] 松平，上田，大沼，森田: Web コンテンツの分析に基づくオントロジー構築および情報整理の試み，人工知能学会　第 4 回セマンティックウェブとオントロジー研究会資料, 2003

[松平 04] 松平，上田，渕上，大沼，森田: 文書からのキーワード抽出と関連情報の収集，人工知能学会第 5 回セマンティックウェブとオントロジー研究会資料, 2004