人工知能学会研究会資料
SIG-SWO-A403-01

語の共起情報に基づくWeb上からの個人メタデータ抽出

Personal Metadata Extraction from the Web using Word Co-occurrance Information

森純一郎[*1*2] 松尾豊[*2] 石塚満[*1]
Junichiro Mori Yutaka Matsuo Mitsuru Ishizuka

[*1]東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, University of Tokyo

[*2]産業技術総合研究所情報技術研究部門
Information Technology Research Institute, National Institute of Advanced Industrial Science and Technology

Abstract: With the currently growing interest in the Semantic Web and Social Networking, personal metadata is coming to play an important role in the Web. This paper proposes a novel keyword extraction method to extract personal metadata from the Web. The proposed method is based on co-occurrence information of words. Our method extracts relevant keywords depending on the context of a person. Our experimental results show that extracted keywords are useful for personal metadata creation. We also discuss the annotation of personal metadata. Annotated documents have strong potential for the Semantic Web applications. Personal metadata is also useful for Social Networking.

1. はじめに

セマンティックウェブ [2] の流れを受け，Web 上では近年，コンテンツに対するメタデータの付加が行われるようになってきている．特に最近は，Weblog などのツールの普及により，エンドユーザは容易に Web コンテンツ作成が可能になってきており Small contents と呼ばれる多様な情報がメタデータとして流通してきている [21]．Web 上のもう一つの新たな動向はソーシャルネットワークである．ソーシャルネットワークは，実社会での友人や知り合いなどの人間関係を Web 上に取り込んだ実世界志向の Web コミュニティである．ソーシャルネットワークはスケールフリーネットワーク [7] の性質を持ち，ネットワーク分析の観点からも注目されている．Weblog における公私を含む多様な個人情報やソーシャルネットワークにおける人間関係に見るように，個人およびその社会的関係というような，より実社会を反映した情報が Web 上に現れてきている．

Web におけるこのような流れの中で，人および人間関係を表現する語彙や記述のためのフレームワークが近年提案されてきている [28, 4, 5, 17]．FOAF (Fried of a Friend)[4] は人および人間関係を XML，RDF を用いて記述するフレームワークである．FOAF が提供する語彙やその他の RDF 語彙を用いて，ユーザは自分の情報や知り合いの情報を記述しプロフィールのメタデータとして自身のホームページや Weblog に付加する．人間関係を表現する FOAF ファイルは，ソーシャルネットワークにおける個人プロフィールとしても利用できる．

セマンティックウェブ実現の課題の一つに，メタデータのアノテーションがある．アノテーションの半自動化および自動化ツール [12, 8, 3] などにより，徐々にメタデータ化された Web コンテンツが普及し始めてはいるが，現在のところ Web 上のほとんどのコンテンツはメタデータを持たない非構造化データである．既存コンテンツをメタデータ化し利用することが今後，セマンティックウェブ普及のための重要な要因となる．

FOAF のような個人メタデータやソーシャルネットワークにおける個人プロフィールは，多くの場合，各ユーザが自身で作成する．これらのデータは，本人しか知りえない情報を含むため，プライバシーの観点から各人が公開すべき情報を選択しつつ作成するのが一般的である．一方で，それらの情報の中には，すでに Web 上の既存の情報源の中で公になっているものが多く存在する．FOAF の語彙の中には個人の活動を示す “組織” や “プロジェクト” などの属性がある．またソーシャルネットワークサービスの多くの個人プロフィールは “所属組織” や “興味” といった項目がある．仮にある人が研究者だとすると，これらの情報は Web 上の個人ページや組織，学会ページに容易に見つけることができる．また，所属組織やプロジェクトのメンバーページや論文の共著者情報など，その人の研究活動の上での知り合い関係情報も Web ページは含んでいる．さらに最近では，Weblog や Web 日記ツールの普及によりユーザの多様な情報が Web 上に現れてきている．

既存の Web ページに含まれるこれらの潜在的な情報は，FOAF のような個人メタデータの自動アノテーションやソーシャルネットワークへの応用の大きな可能性を含んでいる．しかし，既存の研究においては，これらの情報が十分に活用されることはなかった．個人メタデータ抽出としては，特定の文章からの情報抽出研究がある．例えば，論文からの著者情報の抽出などは自然言語処理や機械学習の手法を用いた多くの研究がなされている [9]．しかし，Web ページのように決まった構造を持たず多様な文章を対象とする場合には，特定の文章に特化した既存の情報抽出手法の利用は難しい．

以上の背景，問題点を踏まえて，本研究では Web からの人および人間関係のキーワード抽出方法を提案する．キーワードの元となる語群は，対象とする人名の検索結果の上位ページに含まれる語を用いる．次に検索エンジンのヒット数に基づく共起情報を利用して語群の各語と人との関連度を計算しスコア付けしてキーワードを抽出する．語群には，一般にさまざまなコンテクストの語が含まれている．仮にある人が研究者かつ芸術家であって，それらの活動に関する多くの文章が Web 上に存在すれば，語群にはおのおのの活動に関する語が混在しているはずである．提案手法では，人および人間関係に関するある特定のコンテクストに関連したキーワードを抽出するため，語群とコンテクストについても語の共起情報を用いて関連を考慮する．このようにして語群の中から，各人およびその人の特定のコンテクストに深く関連した語をキーワードとして抽出する．

連絡先: 森純一郎，東京大学情報理工学系研究科，東京都文京区本郷 7-3-1，03-5841-6755，jmori@miv.t.u-tokyo.ac.jp

本論分の構成は以下のようである．2 章ではキーワード抽出の方法を述べる．3 章では提案手法により得られたキーワードの分析および評価の結果を示す．4 章では提案手法の応用に述べる．5 章では関連研究を述べ，6 章では今後の課題を述べる．最後に 7 章においてまとめを行う．

2. キーワード抽出

2.1 キーワード抽出対象

提案手法は，人や人間関係を記述するための語彙で記述されるような属性情報を各人のキーワードとして抽出するものである．例として個人情報を記述する語彙の一つである vCard[28] では (住所，通称，名前，ニックネーム，会社名，役職，職業，電話番号，電子メール，etc) などの属性が定義されている．また，FOAF[4] ではさらに “プロジェクト” や “興味” などの属性や人間関係を表す “知り合い”(knows) という属性が存在する．FOAF において人間関係は knows よって単純化されているが, 人間関係を (友達，同僚，両親，子供，知り合い，etc) などのようにさらに詳細化した語彙も提案されている [5].

提案手法は，このような人および人間関係の属性情報を対象とするものであるが，Web 上に存在する情報は各人によって偏りがある．そこで本研究では対象として特に研究者の情報に焦点をあてる．一般の人に比べると研究者の情報は個人ページ，所属組織および学会ページ，プロジェクトページなどさまざまな Web ページに多くの情報が存在しており，そこには多様な人および人間関係の情報が存在する．

2.2 キーワード候補語抽出

キーワード抽出にあたってまずは，人および人間関係のキーワードの候補となる語群の抽出を行う．本手法では，検索エンジンを用いてある人の氏名を検索したときの検索結果の上位ページに含まれる語をその人のキーワードの候補語として用いる．実際の語群取得に当たっては，まず検索エンジンとして Google[*1] を使用し，氏名の検索結果の上位 10 件の Web ページを取得する．ここで検索結果からさらにリンクはたどらずに，検索結果内の Web ページのみを取得対象とする．取得した Web ページの内で.pdf，.ppt などの html ファイル以外のファイルを除き，html タグの除去などの処理を行う．さらに形態素解析を行い，その結果に対して専門用語抽出ツールである Termex[26] を用いて用語を抽出する．

2.3 語の共起に基づくスコア付け

図 1 にキーワード抽出の手順を示す．氏名の検索結果から抽出されたキーワードの候補語に対して人と語との関連度を考慮したスコア付けを行い最終的なキーワードを決定する．提案手法では，人 (氏名) と語の関連度の尺度として共起情報を用いる．ここで共起とは氏名と語が同一の Web ページに同時に現れること指す．そのような Web ページが多くあるほど両者の関係は強く，語はその人に関連していると考えられる．

実際の共起情報の取得にあたっては，検索エンジンのヒット数を利用する．ヒット数を利用した共起計算の単純な方法としては，氏名と語で AND 検索を行い，そのヒット数を共起とするものである．ここで氏名 n を含む Web ページ集合を N, 語 w を含む Web ページ集合を W とした時にヒット数に基づくこの共起は n および w を含む Web ページ集合から $|N \cap W|$ として与えられる．この共起尺度に基づき提案手法では，Jaccard 係数を用いて氏名を含む Web ページと語を含む Web ページのそれぞれの集合の重なりを考慮した共起を用

*1 http://www.google.co.jp

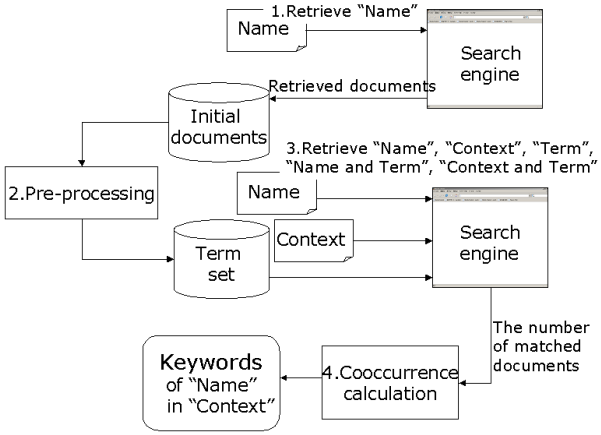

図 1: キーワード抽出の流れ

いる．氏名 n と語 w の単独のヒット数を $|N|$, $|W|$ とした時，Jaccard 係数 $J(n, w)$ は次のように計算できる．

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

このほかの共起の指標として，相互情報量: $log\frac{|N \cup W|}{|N||W|}$，Dice 係数: $\frac{2|N \cap W|}{|N|+|W|}$，Simpson 係数: $\frac{|N \cup W|}{min(|N|,|W|)}$ などの利用が考えられる．

2.4 コンテクストに基づくキーワード抽出

キーワード抽出の基本的な考え方は，キーワード候補語群の中で各人の氏名との共起が大きい語を先の Jaccard 係数に基づいて選ぶものである．しかし，各人のキーワード候補語群はその人の氏名の検索結果中の Web ページから取得したものであるため，語群はその人の異なる複数のコンテクストにまたがった多種多様な語を含んでいる．仮にある人が人工知能，哲学，ロボットの研究をしているとすると語群には，その人のそれぞれの研究活動に関連した語が含まれている．今，その人の人工知能に関するキーワードを知りたいとすると，単に氏名と語の共起にもとづいてキーワード抽出するだけでは哲学やロボットなどのその他の活動の語が混じってしまうために人工知能のキーワードのみを取り出すことはできない．そこで，提案手法では語と氏名の共起に加えて語とコンテクストの共起についても考慮することで，人の特定のコンテクストに関連したキーワードを抽出する．ここでコンテクストとは，観点と捉えることができ，どのような観点からその人のキーワードを抽出したいかに応じてよってコンテクストを決定する．研究者であればコンテクストとして技術用語，研究分野，組織や学会名などを利用することで，これらの特定の観点から研究者のキーワードを捉えることができる．先の例において，コンテクストを人工知能，哲学，ロボットとそれぞれ指定することで各分野ごとの特徴的な語をキーワードとして抽出する．

氏名 n と語 w の共起と同様にコンテクスト c と w の共起指標として Jaccard 係数 $J(c, w)$ を用いると，コンテクストを考慮した最終的なキーワードのスコア $Sscore(n, c, w)$ は，次のように与えられる．

$$Sscore(n, c, w) = J(n, w) + \alpha\ J(c, w)$$
$$(J(n, w) > threshold)$$

これはコンテクスト c における，ある人 n と語 w の関連度を与えるものであり，$Sscore(n,c,w)$ が高いほど語 w は人 n と関連の深いキーワードとなる．α は n と c の関連度を示すパラメータであり $J(n,c)$ などで与えられる．また，$Sscore(n,c,w)$ が高くとも $J(n,w)$ がある閾値 $threshold$ 以上でなければその w をキーワードから除外する．この他に $|W|$ や $|N \cup W|$ などのヒット数をパラメータとし，語のヒット数が閾値以下の場合はキーワードの対象から外している．現在のところこの閾値はヒューリスティックに与えている．

共起を用いたキーワード抽出は，人のキーワードだけでなく人間関係のキーワード抽出にも適用できる．2 人の人 $n1$，$n2$ においてお互いの語群の内でどちらの氏名とも共起する語 w を 2 人の関係のキーワードとすると，キーワードのスコアは $Sscore(n1,n2,w)$ によって決定される．さらに人間関係においてもコンテクスト c を考慮することで 2 者間のキーワードのスコア $Rscore(n1,n2,c,w)$ は次のように与えられる．

$$Rscore(n1,n2,c,w) = Sscore(n1,n2,w) + \beta\ J(c,w) \\ (J(n1,W), J(N2,W) > threshold)$$

これは c における 2 者 $n1$ と $n2$ の関係と語 W の関連度を与えるものであり，$Rscore(n1,n2,c,w)$ が高いほど語 w は 2 者の関係をよく表すキーワードとなる．β は $n1$，$n2$ と c の関連度を示すパラメータであり $J(n1n2,c)$ などで与えられる．また，人のキーワード抽出と同様に $J(n1,w)$，$J(n2,w)$，$|W|$，$|N1 \cup W|$，$|N2 \cup W|$ などのパラメータがある閾値以下の語はキーワードの対象から除外する．

3. 実験

3.1 キーワード分析

提案手法により人および人間関係に関するどのようなキーワードが抽出できるのかを調べるため，FOAF ワークショップ[*2]の 25 名のプログラム委員メンバーおよびセマンティックアノテーションワークショップ (SemAnnot)[*3]の 28 名のプログラム委員メンバーそれぞれについてキーワード抽出を行った [19, 20]．コンテクストはそれぞれのワークショップについて “FOAF”，“Semantic Web” とした．

抽出された各人の上位キーワード (一人あたり約 55 から 60 語，FOAF Workshop では計 1392 語，SemAnnot では計 1646 語) について以下の属性ラベルを手動で付加した．カッコ内は属性ラベルに対応する FOAF 語彙を示している．

- N: 人名 (name,knows)
- T: 専門的な用語 (interest, topicinterest, made, etc)
- E: 学会などのイベント
- P: プロジェクト (currentProject, pastProject)
- O: 組織 (workplaceHomepage, workInfoHomepage)
- URL: (homepage, seeAlso)
- その他: 職業，コミュニティなど

∗2 http://www.w3.org/2001/sw/Europe/events/foaf-galway
∗3 http://km.aifb.uni-karlsruhe.de/ws/semannot2004

これらの属性ラベルを用いて各ワークショップについてそれぞれ表 1 および表 2 に示されるような上位キーワードの属性ラベルの分布を得た．表の中列が示すように，上位キーワードの約半数は用語で占められている．しかし，人名や組織名やプロジェクト名などの他の属性もキーワードとして抽出できていることがわかる．表の右列は各属性がどの程度上位キーワードに含まれているかの一人あたりの平均値であり，人名に関しては約 20 個程度が，組織名やプロジェクト名に関してはから 1 から 3 個が各人の上位キーワード中に含まれていることを示している．この値はあくまで平均値でありキーワードに含まれる属性種は個人差があるが，提案手法がこれらの属性で表現される FOAF のような個人メタデータの抽出に利用できることを示している．これに関連して提案手法の個人メタデータのアノテーションへの応用ついて 5 章で後述する．

属性ラベルの分布に関して 2 つのワークショップの間には類似性があるが，URL やイベント属性など差異も見られる．これらの差異はキーワード候補語を抽出する際に使用する Web ページの種類に起因するものである．表 4，5 はキーワード候補語を抽出する際に使用した Web ページ種の分布を示している．FOAF ワークショップはワークショップの性質上，Weblog やソーシャルネットワークを積極的に利用しているメンバーが多く，Web ページにはそれらのページが多く含まれている．FOAF ワークショップメンバーのキーワードにて用語や URL など属性が多く含まれるのは Weblog ページに起因している．一方，FOAF ワークショップに比べると SemAnnot ワークショップでは候補語のソースとして学会などのイベントページや論文などが Web ページ種として多く含まれる．そのためにキーワードには人名やイベント名が多く含まれている．このように抽出されるキーワードの属性はキーワード候補語が含まれる Web ページの種類に関連している．今後はもととなる Web ページ種と抽出されるキーワードの属性の関係を詳細に調べる必要がある．

3.2 キーワード評価

提案手法のキーワード抽出の精度に関して precision および coverage による評価を行った．また，あわせて提案であるコンテクストの概念の有効性についても precision により評価を行った．実験は，人工知能の研究を行っている 6 人の研究者に対して行い，それぞれの被験者についてその氏名を検索して得られた検索結果上位 10 件のドキュメント (html のみを対象) に対して tf(Term Frequency)，tf-idf(Term Frequency-Inverse Document Frequency)，コンテクストを含まない氏名と語の共起に基づく手法，提案手法の 4 つの手法を用いてキーワードを抽出し比較を行った．

tf-idf の計算に当たっては 2004 年度人工知能学会の参加者 567 名の氏名を検索して得られた計 3981 個の html ファイル (一人最大 10 個) をコーパスとした．また，語 w に対する idf の重み付けは $log(D/df(w)) + 1$ とした．ここで D はコーパスの全ドキュメント数，$df(w)$ はコーパスの中で語 w が出現するドキュメント数である．各人の名前を検索して得られた html ファイルそれぞれについて tf-idf を計算し正規化をおこなった後で tf-idf 上位の語をキーワードとして抽出した．共起を用いた手法に関して，氏名と語の共起の計算には Jaccard 係数を用いた．提案手法におけるコンテクストには “人工知能” を使用した．

各被験者について上記 4 つの各手法を用いてそれぞれ上位 20 個のキーワードを抽出し，手法間で重複するキーワードを除いた後で各手法から得られたキーワードを混合した．抽出されたキーワードについて，各被験者に Q1.「自身の研究活動に

表 1: 上位キーワードに付加された属性ラベルの分布 (FOAF ワークショップメンバー)

属性ラベル	総数	一人平均
用語	695 (49.9%)	27.8
人名	476 (34.1%)	19.04
組織	71 (5.1%)	2.84
URL	65 (4.6%)	2.6
プロジェクト	35 (2.5%)	1.4
イベント	31 (2.2%)	1.24
その他	19 (1.3%)	0.76
計	1392	

表 2: 上位キーワードに付加された属性ラベルの分布 (SemAnnot ワークショップメンバー)

属性ラベル	総数	一人平均
人名	767 (46.5%)	27.3
用語	613 (37.2%)	21.8
イベント	105 (6.3%)	3.7
組織	73 (4.3%)	2.6
プロジェクト	48 (2.5%)	1.7
URL	40 (2.4%)	1.4
その他	0	0
Total	1646	

表 3: キーワード候補語を含む Web ページ種 (FOAF ワークショップメンバー)

Web ページ種	総数
個人ページ	58 (23.3%)
Weblog	465 (19.2%)
記事	25 (10.0%)
イベントページ	14 (5.6%)
ML ログ	14 (5.6%)
論文	13 (5.2%)
DBLP	12 (4.8%)
組織ページ	10 (4.0%)
書籍ページ	10 (4.0%)
ソーシャルネットワーク	9 (3.6%)
オンラインコミュニティ	6 (2.4%)
プロジェクトページ	5 (2.0%)
その他	39 (15.6%)
Total	250

表 4: キーワード候補語を含む Web ページ種 (SemAnnot ワークショップメンバー)

Web ページ種	総数
個人ページ	73 (26.0%)
イベントページ	32 (11.4%)
ML ログ	27 (9.6%)
論文	26 (9.2%)
DBLP	22 (7.8%)
組織ページ	17 (6.0%)
プロジェクトページ	16 (5.7%)
書籍ページ	11 (3.9%)
Publication リスト	8 (2.8%)
Weblog	6 (2.1%)
その他	42 (15.0%)
Total	280

関連する語をチェックして下さい」という質問を行った．これより，各手法で抽出された 20 個の中にユーザがチェックをした語が含まれる割合をその手法の precision として評価する．また，Q2.「Q1 でチェックした語の中で自身の研究活動を表すのに不可欠な語を 5 つチェックしてください」という指示を行った．これにより，被験者が選んだ語の中に各手法により抽出された語が含まれる割合を，その手法の coverage として評価する．最後に，Q3.「Q1 でチェックした語の中で特に人工知能分野という観点から自身の研究活動に関連すると思う語をチェックしてください」という質問を行った．これにより，各手法で抽出された 20 個の中にユーザがチェックをした語が含まれる割合をコンテクストに基づく precision(context precision) として，コンテクストに関連したキーワードが抽出できているかを評価する．

表 5 は全被験者に対する各手法の precision，coverage および context precision を示している．precision，coverage ともに提案手法が，他の手法を上回っている．tf，tfidf は語の出現頻度に基づいてキーワード抽出するために一般語が含まれてしまうのに対し，共起および提案手法では氏名と語の共起を用いることで各人に関連した語を抽出できていることがわかる．また，context precision が示すように氏名と語の共起に比べて，さらにコンテクストと語の共起を考慮した提案手法のほうがコンテクスト (“人工知能”) に応じた各人のキーワードを抽出できていることがわかる．このことは，多様な情報から特定の文脈に応じた情報を抽出する際に，我々の提案するコンテクストの概念が有効であることを示している．

表 5: 各手法のの precision，coverage，context precision(被験者 6 名)

手法	tf	tfidf	共起	提案
precision	9.17%	15.00%	51.09%	**53.26**%
coverage	16.67%	20.00%	46.67%	**53.33**%
context precision	5.00%	3.33%	18.48%	**22.82**%

4. 応用

4.1 ソーシャルネットワーク

人および人間関係のキーワードを抽出する提案手法はソーシャルネットワークへ応用できる．松尾らは人間関係を利用した情報支援を目的として Web 上から研究者間の協働関係などを抽出する手法を提案している [16]．図 2 は Web から抽出された人工知能学会メンバーのソーシャルネットワークを示している．

表 6，7 は抽出した人および人間関係のキーワードの例である．抽出にあたって使用したコンテクストは “人工知能” である．キーワードとして組織名や関係者名や研究に関する用語などコンテクストに関連した各人および関係の多様なキーワードが抽出できていることがわかる．抽出されたキーワードを用いてソーシャルネットワーク上で関連する人の検索や 2 者にどのような関係があるのがチェックするなどキーワードは，ソーシャルネットワーク上に意味を与えた上でさまざなサービスを可能にする．

表 6: キーワードの例:“松尾豊”，“石塚満” の上位キーワード

松尾豊	石塚満
人工知能	石塚満
産業技術総合研究所–	人工知能
–サイバーアシスト研究センター	東京大学大学院–
産業技術総合研究所–	–情報理工学系研究科
–知能システム研究部門	東京大学教授
松尾豊	岡崎直観
友部博教	土肥浩
人工知能学会全国大会	松尾豊
スケジューリング支援システム	人工知能基礎論
人間関係ネットワーク	マルチモーダル擬人化インタフェース
イベント空間情報支援プロジェクト	知能的 WWW 情報空間
人工知能学会	MPML
ユーザ行動モデル構築	擬人化エージェント
イベント空間支援	キャラクタエージェント
濱崎雅弘	大澤幸生
宮崎伸夫	相澤清晴
中村嘉志	松村真宏
石塚満	東京大学石塚研究室

表 7: “松尾豊”，“石塚満” の関係キーワード

松尾豊–石塚満
石塚満
松尾豊
産業技術総合研究所-
サイバーアシスト研究センター
人工知能
石塚研究室
岡崎直観
土肥浩
大澤幸生
橋田浩一
松村真宏
人間関係ネットワーク

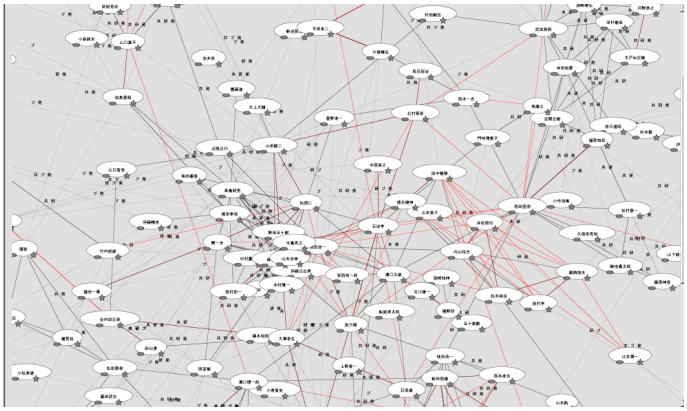

図 2: Web から抽出した人工知能学会研究者間のソーシャルネットワーク

4.2 セマンティックアノテーション

3 章で示したように，提案手法を用いて人や人間関係のさまざまな属性が抽出できる．抽出したキーワードと属性を適切に対応づけ，アノテーションツール [12, 8, 3] を使用することによりキーワードが含まれるもとの Web ページに対して，メタデータを付加することが可能である．また，抽出したキーワードを用いて図 3 に示す FOAF ファイルのような個人メタデータを半自動生成し，対話的にユーザが適宜修正，追加を行うことによりユーザは容易に自身のメタデータを作成できる．このように提案手法はメタデータのアノテーションにおいて多くの利用可能性を持っている．

5. 関連研究

本研究は情報抽出，特に人や人間関係に関連した情報の抽出である．情報抽出の対象として本研究は Web を対象としている．人に関する情報抽出の内で対象文章を Web としないものとしては，論文データベースからの著者の所属抽出 [9] や社内の業務文章からの従業員情報抽出 [11] がある．これらの手法は，事前に定義された文章形式やヒューリスティックルールおよび特定のエンティティに関する機械学習といったものを使用している．さらに抽出すべき固有表現，オントロジー，ドメイ

図 3: 抽出されたキーワード (表 6) を利用した FOAF ファイルの例

```
<foaf:Person>
<foaf:mbox rdf:resource=""/>
<foaf:name>石塚満 </foaf:name>
<foaf:interest
rdfs:label="擬人化エージェント" rdf:resource=""/>
<foaf:currentProject
rdfs:label ="マルチモーダル擬人化インタフェース"
rdf:resource=""/>
<foaf:workplaceHomepage
rdfs:label="東京大学" rdf:resource=""/>
<foaf:knows>
<foaf:Person>
<foaf:mbox rdf:resource=""/>
<foaf:name>松尾豊</foaf:name>
......
<rdfs:seeAlso rdf:resource= "page contains many
keywords">
```

ンなどをあらかじめ限定している．しかしデータベースや社内文章のような構造化された情報源に対して，Web 上の情報は一般に非定型であり，多様性を持っているため抽出にあたって事前になんらかの前提を与えることは難しい．また，従来研究において人に関する情報抽出は氏名，所属情報，メールアドレスなど限定的な情報抽出にとどまっていた．しかし，Web には多岐に渡る情報が存在する．我々の手法は Web などの構造化されてない情報源からの人に関する多様な情報抽出を可能にするものである．

Web 上からの人の関する情報抽出研究として，Web ページから人名を収集し人間関係ネットワークを構築し，かつそのネットワーク上で特定の専門用語と関連する人物の検索を行う Referral web[13] がある．同様なシステムに検索語と関連する人物を Web から発見する NEXAS,KeyPerson[10] がある．どちらの研究も検索エンジンの結果を利用し，また共起を用いて語と人の関連性を調べている点は本研究と類似している．しかし、人と語の関連において両研究ともに，ある語に主眼を置き，それに関連する人を Web から抽出するというアプローチをとっているのに対し，我々の主眼は人や人間関係にあり，ある人のキーワードを Web から抽出することで応用として人のメタデータの自動生成やソーシャルネットワークへの利用を考慮している．その点で我々が用いている人のコンテクストといった概念は新規性が高い．

山本らは政治家などの職業名を入力として，検索エンジンとハイパーリンクを利用して，特定の職業の人物情報を網羅的に収集する方法を提案している [23]．方法はターゲットとなる職業に関して，表形式で書かれた人名録が存在することを前提にしており，Web ページの構造に依存した限定的な手法である．一方で我々の手法は Web ページの構造によらず、かつ任意の人名を扱うことが可能なキーワード抽出手法である．

検索エンジンを利用せずに人に関する情報を抽出研究もある．松平らは，Web やイントラネットの上の情報源から，あらかじめ定義されたオントロジーに対応したヒューリスティックルールを用いて技術情報や人に関する情報抽出研究を行っている [18]．同様な研究として，Alani らは芸術家についてのバイオグラフィー情報を Web から抽出する研究を行っている [1]．彼らの手法はあらかじめ定義した (主語-関係-オブジェクト) という語彙的連鎖関係およびオントロジーを用いて情報を抽出するものである．しかし，Web ページはしばしば定型的な記述を含まず語彙的関係やヒューリスティックルール適用するのはしばしば困難である．Dingli らは，大学研究者の名前，プロジェクト，発表文献といった情報を教師なし学習を用いて抽出する研究を行っている [6]．情報抽出にあたっては”seed” 情報として事前にユーザによって提供された情報をもとに学習を行う．これらの研究が事前に定義されたオントロジーやユーザから提供される情報が必要なのに対して，我々の手法は名前のリストのみで人および人間関係に関連した多様な情報が抽出可能である．

Velardi らは，あるドメインについてキーワードを Web から抽出する研究を行っている [22]．彼らの手法は，Web ページ内および Web ページ群の語の出現頻度に基づいたものである．同様にキーワード抽出には tfidf のように文章コーパスを利用した語の出現頻度に基づく手法が用いられるが，我々の手法はコーパスを必要とせず検索エンジンのヒット数を利用した語の共起情報のみによりキーワード抽出が可能である．また，特定ドメインに限定せずに使えることも特徴である．

情報抽出からの観点に加えて，本研究は Web を利用した質問応答システム [14] と捉えることもできる．しかし，我々のアプローチは正解集合を探すというよりは，ある人や人間関係に関連した情報を Web を用いて幅広く抽出するものである．

6. 課題

2 章で述べたように，現在のところ提案手法は研究者情報を対象としている．将来的には，研究者情報に特化したものではなく，一般的な人や人間関係のキーワード抽出手法へ拡張すること考えている．特に最近では Weblog などのさまざまな個人コンテンツ作成ツールにより，今後ますます Web 上に個人に関連した情報が増加すると予想される．そのような情報が増加すれば，提案手法は適用範囲が広がると考える．

キーワード抽出にあたっていくつかの課題が存在する．キーワード候補語を取得するために，ある氏名を検索する際に同姓同名が存在する場合，その検索結果には同姓同名の複数の人に関係した Web ページが含まれる．現在のところ，これを解決するために氏名に加えて所属組織名を検索語に加えることで検索結果から同姓同名をなるべく除くようにしている．研究者などの場合は，学会の参加情報や論文の著者情報により，容易に所属情報が取得できるが，一般には必ずしも人の所属情報が利用できるとは限らない．さらに所属情報を付加することはキーワードの計算以前に，コンテクストを限定することになる．今後は同姓同名に対する一般的な解決手法を検討する必要がある．

キーワード候補語の取得にあたり検索結果の内で上位何件の Web ページを対象とするか，また検索結果からたどるリンクの深さについては検討する必要がある．提案手法は語ごとに検索エンジンのヒット数を用いて共起を計算するため，キーワード候補語群が大きくなるほどキーワードを計算するためのクエリー数が増えキーワード抽出に時間を要する．上位 10 件から候補語を取得してもその候補語数は一人当たり 1000 語程度になり，さらに候補語が増えれば語と氏名および語とコンテクストのヒット数を得るには膨大なクエリーが必要となる．そのため計算時間および検索エンジンへの負荷を考慮して，取得対象の Web ページを現在のところ各人につき 10 件とした．しかし，上位 10 件の Web ページに必ずしも適当な語が含まれているとは限らないため，キーワード候補語の取得対象として検索結果内でどのような Web ページを使用するのがよいのか今後分析を行う必要がある．また，将来は html ファイルに限らずその他のファイル種も語群の取得対象として利用する予定である．

共起の計算には現在のところ Jaccard 係数を用いている．しかし，他にもさまざまな共起の指標があるためキーワード抽出に最適な共起の計算方法を現在検討中である．共起計算の洗練に加えて，今後は応用で述べたように提案手法をメタデータのアノテーションへ利用するために，属性ラベルの定義やキーワードの属性の自動判定などを行う予定である．

7. まとめ

本論文では，語の共起情報を用いて Web 上から人および人間関係のキーワードを抽出する手法を提案した．実験を通して提案手法により人および人間関係に関するさまざまな属性が取得できることを示した．Weblog などの普及により，Web 上には今後ユーザに関するさまざまな情報が流通するだろう．そのような中で多様な情報の中から，ユーザに関する適切な情報を抽出する本手法はソーシャルネットワークやセマンティックアノテーションなどセマンティックウェブの観点から今後重要な役割を果たすと考える．

参考文献

[1] H. Alani et al. Automatic Extraction of Knowledge from Web Documents. *In Workshop of Human Language Technology for the Semantic Web and Web Services, 2nd International Semantic Web Conference*, Sanibel Island, Florida, USA, (2003).

[2] T. Berners-Lee, J. Hender, O. Lassila. The Semantic Web. Scientic American, (2001).

[3] F. Ciravegna, A. Dingli, D. Petrelli, and Y. Wilks. User-system cooperation in document annotation based on information extraction. *In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management, EKAW02*, Springer Verlag, (2002).

[4] Dan Brickley and Libby Miller. FOAF: the 'friend of a friend' vocabulary. http://xmlns. com/foaf/0.1/, (2004).

[5] Ian Davis and Eric Vitiello Jr. RELATIONSHIP: A vocabulary for describing relationships between people, http://vocab.org/relationship/, (2004).

[6] A. Dingli, F. Ciravegna, D. Guthrie, Y. Wilks. Mining Web Sites Using Usupervised Adaptive Information Extraction. *In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, (2003).

[7] L. Garton, C. Haythornthwaite, and B. Wellman. Studying online social networks. In Doing Internet Research, S. Jones, Ed. Sage, Thousand Oaks, CA, pp. 75–105 (1999).

[8] S. Handschuh, S. Staab, and F. Ciravegna. S-CREAM–Semi-Automatic Creation of Metadata. *Semantic Authoring, Annotation and Markup Workshop*, France, (2002).

[9] Hui Han, et al. Automatic Document Metadata Extraction using Support Vector Machines. In Proceedings of the *ACM IEEE Joint Conference on Digital Libraries*, 37-48 (2003).

[10] 原田，佐藤，風間：Web 上のキーパーソンの発見と関係の可視化，情報処理学会研究報告， DBS-130/FI-71， (2003).

[11] 井形，小櫻，片山，津田：セマンティックグループウェア：RDF を用いた Knowwho の実現，セマンティックウェブとオントロジー研究会，A303-05 (2004).

[12] J. Kahan and M. R. Koivunen. Annotea: An open rdf infrastructure for shared web annotations. *In Proceedingsof the 10th International WWW Conference*, pp.623–632 (2001).

[13] H. Kautz, B. Selman, M. Shah. The Hidden Web. *AI Magazine*, Vol.18, No.2, pp.27–36 (1997).

[14] C. C. T. Kwok, O. Etzioni, D. S. Weld.Scaling question answering to the web, *In Proceedings of the 10th International Conference on WWW*, pp.150–161 (2001).

[15] Lars Marius Garshol. Living with topic maps and RDF. http://www.ontopia.net/topicmaps/materials/tmrdf.html, (2003).

[16] Y. Matsuo, H. Tomobe, K. Hasida, M. Ishizuka. Mining Social Network of Conference Participants from the Web. *In Proceedings of the International Conference on Web Intelligence*, pp.190–194 (2003).

[17] Y. Matsuo, M. Hamasaki, J. Mori, H. Takeda and K. Hasida. Ontological Consideration on Human Relationship Vocabulary for FOAF. *In Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and Semantic Web*, (2004).

[18] 松平，上田，大沼，渕上，森田：文章からのキーワード抽出と関連情報の収集，セマンティックウェブとオントロジー研究会，A303-02 (2004)

[19] J. Mori, Y. Matsuo, M. Ishizuka, and B. Faltings. Keyword Extraction from the Web for FOAF Metadata. *In Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and Semantic Web*, (2004).

[20] J. Mori, Y. Matsuo, M. Ishizuka, and B. Faltings. Keyword Extraction from the Web for Personal Metadata Annotation. *In Proceedings of the 4rd International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot 2004)*, (2004).

[21] 武田英明：Weblog 研究の現状，セマンティックウェブとオントロジー研究会，A402-06 (2004).

[22] P. Velardi, M. Missikoff, R. Basili. Identification of relevant terms to support the construction of Domain Ontogies. *In ACL-EACL Workshop on Human Language Technologies*, Toulouse, France, (2001).

[23] 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，2000-ICS-119-24，pp.173–180 (2000).

[24] R. Yangarber and R. Grishman. Machine Learning of Extraction Patterns from Unanotated Corpora: Position Statement. *Workshop Machine Learning for Information Extraction*, IOS Press, Amsterdam, pp.76–83 (2000).

[25] Resource Description Framework(rdf) Schema Specification. In *W3C Recommendation*, (2000).

[26] http://gensen.dl.itc.u-tokyo.ac.jp/win.html

[27] DAML Ontology Library. http://www.daml.org/ontologies/

[28] Representing vCard Objects in RDF/XML. http://www.w3.org/TR/2001/NOTE-vcard-rdf-20010222/