論 文

Technical Papers

Webからの人物に関するキーワード抽出

Personal Keyword Extraction from the Web

森 純一郎 Junichiro Mori
東京大学情報理工学系研究科
School of Information Science and Technology Engineering, University of Tokyo
jmori@miv.t.u-tokyo.ac.jp, http://www.miv.t.u-tokyo.ac.jp/~jmori/

松尾 豊 Yutaka Matsuo
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
y.matsuo@carc.aist.go.jp, http://carc.aist.go.jp/~y.matsuo/

石塚 満 Mitsuru Ishizuka
東京大学情報理工学系研究科
School of Information Science and Technology Engineering, University of Tokyo
ishizuka@miv.t.u-tokyo.ac.jp, http://www.miv.t.u-tokyo.ac.jp/~ishizuka/

keywords: keyword extraction, word co-occurrence, search engine, metadata, social network

Summary

With the currently growing interest in the Semantic Web, personal metadata to model a user and the relationship between users is coming to play an important role in the Web. This paper proposes a novel keyword extraction method to extract personal information from the Web. The proposed method uses the Web as a large corpus to obtain co-occurrence information of words. Using the co-occurrence information, our method extracts relevant keywords depending on the context of a person. Our evaluation shows better performance to other keyword extraction methods. We also discuss the lessons that is learned from a preliminary experiment.

1. ま え が き

近年，WeblogやWikiなどのツールの普及により，ユーザは容易にWebコンテンツを作成することが可能になってきている．これらのツールでは，Web上の操作だけで容易にWebページを更新することができ，情報公開やコミュニケーションのためのさまざまな機能が利用できる．それに伴いSmall contentsと呼ばれる多様な情報が流通している [武田 04]．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サービス (SNS) である．SNSはすでに全世界に数百以上存在し，中には数百万を超えるユーザが参加するものもある．Small contentsにおける公私を含む多様な個人情報やSNSにおける人間関係に見るように，実社会での友人や知り合いとのコミュニケーションをWeb上に取り込んだ実世界志向のコミュニティがWeb上に現れている．

セマンティックウェブの枠組みの中では，個人の情報を公開するための様々なオントロジーが提案されている．例えば，個人情報を記述する語彙であるvCard [vca] 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，電子メールなどの属性が定義されている．また，FOAF (Friend of a Fined) [Brickley 04] では参加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)」などの多様な属性を定義している．FOAFにおいて人間関係はknowsに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，知り合いなどさらに詳細化した語彙も提案されている [Davis 04, Matsuo 04]．セマンティックウェブで現在，ユーザに最も多く利用されているオントロジーはFOAFであり [Ding 05]，個人や人間関係の情報を記述することに対する関心が高まっている．

一方，近年では研究者に関する多様な情報が，研究者個人のページ，組織のページや研究プロジェクトページなど多くのWebページで公開されている．研究者に関する情報を適切に集め，学際的な研究および産学連携に生かす試みも行われている [松尾 05]．情報系の研究者は特に，最も情報化の早いユーザであると考えられるが，研究者に関する情報を自動的に抽出し，研究活動の情報を公開するために利用することができれば，個人情報の自動取得の一例を示すことができるだろう．

本論文では，Web上の情報を用いて，研究者の情報をキーワードとして自動的に抽出する手法を提案する．キーワードとは，研究者の活動を表すために重要な語であり，例えば，所属組織，研究テーマ，共著者，プロジェクト名である．これらはFOAF [Brickley 04] の語彙のなかでは `foaf:currentorganization`，`foaf:interest`，

foaf:knows，foaf:currentproject などにあたる属性を自動的に抽出することに相当する．キーワードの抽出にあたっては Web ページにおける語の共起に着目し，検索エンジンを用いて共起情報を取得する．語の共起情報を統計的に処理することで，研究者のキーワードを抽出する．

本論文の構成は以下である．2 章ではキーワード抽出の手法を述べる．3 章では評価を行う．4 章では考察と議論を行う．5 章では関連研究を述べ，最後に 6 章においてまとめを行う．

2. Web 上の情報からのキーワード抽出

2·1 キーワード抽出の従来手法と Web への適用可能性

キーワード抽出[*1]は，文書において重要な語を抽出する技術であり，情報検索，要約，質問応答など，自然言語処理の幅広い分野で用いられている．

キーワード抽出の際の語の重み付けには，網羅性と特定性という 2 つの性質を考慮する必要がある [徳永 98]．すなわち，キーワード抽出の一般的な指針としては網羅性と特定性を兼ね備えた語の重要度が上がるように重みを与えるべきである．現在，多くの自然言語文書処理において用いられている $tfidf$ (Term Frequency-Inverse Document Frequency) は，対象とする文書に頻繁に出現し (網羅性)，コーパス中のほかの文書にあまり出現しない (特定性) ような語を重要と見なすもので，一般的には次のような式で与えられる．

$$tfidf(t,d) = tf(t,d) \cdot idf(t)$$

これはある文書 d における語 t についての $tfidf$ 値を表している．$tf(t,d)$ は語の出現頻度，$idf(t)$ は語がコーパスの全文書集合のどれぐらいの文書に出現するかの尺度であり $log(|D|/df(t)+1)$ などで表される．ここで $|D|$ はコーパスの全文書数，$df(t)$ は語 t が出現するコーパスの文書数である．idf は語の情報量の単純で頑健な推定値となっており，tf を出現確率とみなすと $tfidf$ は出現確率と情報量をかけあわせた特徴量となっている [相澤 00]．$tfidf$ の他にも，文書集合中の語の分布をもとに，統計的，経験的な尺度を用いたさまざまなキーワード抽出法が提案されている [大澤 99, 松尾 02]．

さて，Web 上の情報からある人物に関連するキーワードを抽出することを考えよう．ここでは，研究者を対象とする．まず考えられるのが，対象を検索し，得られた Web ページからキーワード抽出する方法である．例えば，以下のような処理である．

（1）研究者の氏名で Web ページの検索を行う．
（2）検索にヒットした上位のページを取得する．

*1 重要語抽出，語の重み付けとも同様の技術である．また，情報検索においては索引語の抽出と同様である．

表 1 $tfidf$ により抽出した「石塚満」の上位キーワード

東京大学
JAVA アプリケーション
キャラクタ
シナリオ創発表
研究科
電子
マイクロソフト
伊庭研
大澤幸生
研究成果
工学

（3）その各ページを対象に，他の多くの Web ページをコーパスとして，$tfidf$ 値の高い語をキーワードとして抽出する．

表 1 に，この方法で抽出した「石塚満」のキーワードを示す．コーパスは， 2004 年度人工知能学会全国大会の参加者 567 名の氏名を検索して得られた計 3981 個の html ファイル (1 人につき最大 10 個) を使用した．抽出されたキーワードを見ると,「キャラクタ」や「研究成果」といった研究に関する語も含まれるが，具体的な研究内容を表すような語は得られていない．また,「電子」や「工学」のような，キーワードとしては一般的すぎる語が上位にきている．この原因として次のような点が指摘できる．

- Web ページは新聞記事や論文と比べ，多様性が高く語の出現頻度 (tf) がキーワードのよい目安となりにくい．つまり，研究に関するキーワードが Web ページ中に何個も出現するとは限らない．
- Web に出現する語は多様であり，複数の単語から構成されるフレーズまで含めると，語が出現する適切な文書数 (df) を得られるほど大量の Web ページを集めるのは難しい．

Web における語の多様性を考えれば， 研究に関連する語彙をあらかじめ与えておくことも難しい．また，研究者のページだけを集めることによって，研究に関連する一般語の $tfidf$ 値を下げ，特徴的な語の $tfidf$ 値を大きくすることは可能である．しかし，同時に研究に関連しない趣味やビジネスなどの語も $tfidf$ 値が高くなってしまう．

2·2 提案手法のアイディア

従来のキーワード抽出が主に対象としてきた新聞記事や論文といった文書に比べ，Web ページは内容や形式が多様であり，文書を書く目的も一定でないことから，従来のキーワード抽出手法をそのまま用いることは難しい．本論文では，Web ページの多様性や不均質性に対応し，目的に応じたキーワードを抽出する方法を提案する．

提案手法のアイディアを示すのが図 1 である．本論文では 2 つの語が同じ Web ページ内に出現することを語が共起するといい，図中の語のネットワークは，Web 上における語の共起関係を視覚化したものである．この例では,「松尾豊」という語と,「ソフトボール」「うどん」「人

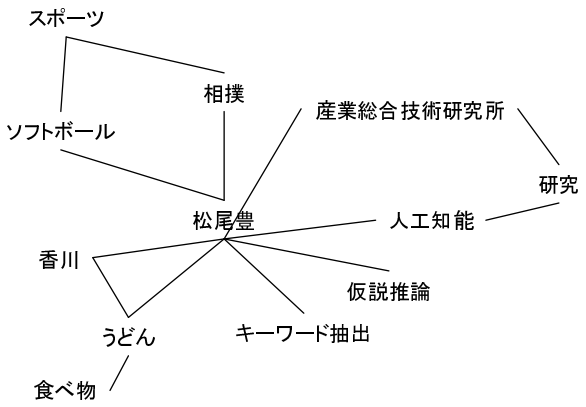


図 1 Web 上の語の共起ネットワークの例

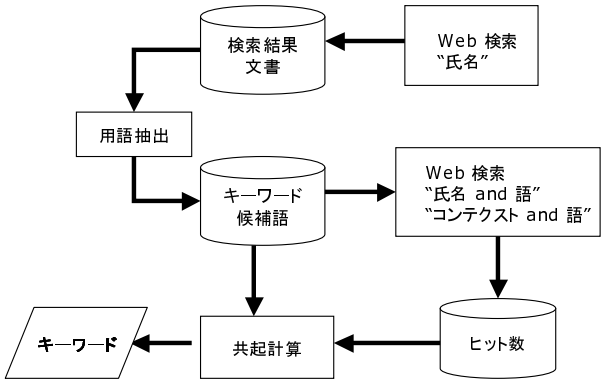


図 2 キーワード抽出の流れ

工知能」「仮説推論」など多くの語が共起している．特に，この中でキーワードになりやすいのは,「松尾豊」という語と強く共起する語，つまり多くの Web ページで同時に出現する語であると考えられる．したがって，提案手法のひとつ目の仮説は以下である．

仮説 1 Web 上で氏名と特に強く共起する語がその人物のキーワードになりやすい．

次に，たとえ氏名と強く共起していても，趣味に関する「ソフトボール」や「うどん」などの語は，研究に関するキーワードとは言いがたい．したがって，キーワードとして選ばれる語は，キーワード抽出の目的と合致しているかことが必要である．本手法では，目的に沿ったキーワード抽出を行うために，**コンテクストワード**と呼ぶ語を定義する．コンテクストワードは，抽出するキーワードの文脈を特定するものであり，文脈に合致しない語を除外する働きがある．例えば,「松尾豊」の研究に関連するキーワードを知りたい場合，コンテクストワードを「研究」とし,「松尾豊」と強く共起する語の中でも，特に「研究」と共起する語だけに絞り込むことにより,「人工知能」という語を抽出することができる．これは，語の共起ネットワーク上で氏名およびコンテクストワードのどちらにも近いような語をキーワードとすることに相当する．コンテクストワードを「食べ物」にすれば,「うどん」が抽出されればよい．したがって，2 つ目の仮説は以下である．

仮説 2 特定の文脈を表すためのコンテクストワードと強く共起する語は，その人物のその文脈におけるキーワードになりやすい．

本論文では，このような 2 つの仮説に基づいたキーワード抽出手法を構築する．なお，これらの仮説は，3 章の評価実験によって検証する．

氏名と語，語とコンテクストワードの共起の強さは，検索エンジンを用いて求めることができる．例えば，氏名と語の AND 検索結果のヒット件数は，Web 上での語の共起の強さを表す指標である．提案手法では語の共起の強さを表す指標として，Jaccard 係数を用いる．例えば,「石塚満」と「高速推論」の共起の強さは,「石塚満」を検索クエリーとしたときの検索ヒット数を $|A|$,「高速推論」を検索クエリーとしたときの検索ヒット数を $|B|$,「石塚満 AND 高速推論」を検索クエリーとしたときの検索ヒット数を $|A \cap B|$ とすると，Jaccard 係数 $|A \cap B|/(|A| + |B| - |A \cap B|)$ で与えられる．これは氏名を含む 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度合いを表すものである．Jaccard 係数が大きいほど，2 つの集合の重なりが大きく，関連が強いことになる．他にも共起の強さを測るものとして，共起頻度，相互情報量，ダイス係数，overlap 係数，余弦などがある [Manning 99] が，Jaccard 係数はネットワークの関係の強さを求めるときにはよく使われる一般的な尺度である [Kautz 97].

2·3 提案手法の詳細

図 2 に提案手法の流れを示す．まず，検索エンジンを利用して人物の検索を行い，検索結果の上位 Web ページを取得する[*2]．取得した Web ページに対して html タグの除去および形態素解析を行い，Termex [ter][*3]を用いて用語を抽出する．

次に Web ページから抽出した語と名前との共起の強さを Jaccard 係数を用いて求める．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 として，n と w の単独のヒット数および n と w の AND 検索のヒット数をそれぞれ $|N|$, $|W|$, $|N \cap W|$ として表す．この時，Jaccard 係数 $J(n,w)$ は次のように計算できる．

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

次にコンテクスワード c と語 w の共起の強さを，同様に求める．c を含む Web ページ集合を C とすると，$J(c,w)$ は次のように計算できる．

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

*2 検索エンジンとして Google(http://www.google.co.jp) を使用した．また，後述の同姓同名問題に対応するために所属情報を検索クエリーに加えた．

*3 東京大学中川研究室，横浜国立大学森研究室で開発された用語抽出システム．

表 2 コンテンツワード「人工知能」に対する「石塚満」の上位キーワードとスコア

$\alpha=0$		$\alpha=r/3$		$\alpha=r$	
岡崎直観	0.081	人工知能学会	0.170	自然言語	0.083
松尾豊	0.069	岡崎直観	0.081	岡崎直観	0.081
土肥浩	0.065	松尾豊	0.070	松尾豊	0.071
キャラクタエージェント	0.065	土肥浩	0.065	人工知能学会誌	0.069
擬人化インタフェース	0.053	キャラクタエージェント	0.065	自然言語処理	0.066
擬人化エージェント	0.046	自然言語	0.055	土肥浩	0.066
東京大学工学部電子情報工学科	0.043	擬人化インタフェース	0.053	キャラクタエージェント	0.065
高速仮説推論	0.043	人工知能学会誌	0.048	擬人化インタフェース	0.053
黒橋禎夫	0.039	擬人化エージェント	0.047	擬人化エージェント	0.048
仮説推論	0.034	自然言語処理	0.044	人工知能学会全国大会	0.047
マルチモーダルメディア	0.034	東京大学工学部電子情報工学科	0.044	知識ベース	0.045
高速仮説推論システム	0.032	高速仮説推論	0.043	マルチエージェント	0.045
谷内田正彦	0.031	黒橋禎夫	0.040	東京大学工学部電子情報工学科	0.044
空間共有コミュニケーション	0.031	仮説推論	0.036	ヒューマンインタフェース	0.044
仮説推論システム	0.031	マルチモーダルメディア	0.034	高速仮説推論	0.044
擬人化エージェント	0.031	谷内田正彦	0.033	機械学習	0.043
感性基盤機能	0.028	高速仮説推論システム	0.033	インタラクション	0.041
相澤清晴	0.027	人工知能学会全国大会	0.032	黒橋禎夫	0.041
キーワード抽出法	0.027	空間共有コミュニケーション	0.031	知識処理	0.039
産業技術総合研究所	0.027	仮説推論システム	0.031	仮説推論	0.038
大澤幸生	0.026	マルチエージェント	0.030	知識表現	0.037

表 3 コンテンツワード「大学」に対する「石塚満」の上位キーワードとスコア

$\alpha=r/3$		$\alpha=r$	
岡崎直観	0.081	工学系研究科	0.083
松尾豊	0.069	岡崎直観	0.081
土肥浩	0.064	人工知能	0.079
キャラクタエージェント	0.062	松尾豊	0.069
擬人化インタフェース	0.054	土肥浩	0.064
工学系研究科	0.050	大学院工学系研究科	0.063
東京大学工学部電子情報工学科	0.043	キャラクタエージェント	0.062
高速仮説推論	0.040	東京大学工学部	0.060
人工知能	0.040	擬人化インタフェース	0.054
擬人化エージェント	0.035	東京大学大学院工学系研究科	0.054
マルチモーダルメディア	0.034	国立情報学研究所	0.053
仮説推論	0.033	奈良先端科学技術大学院	0.052
大学院工学系研究科	0.032	東京大学工学部電子情報工学科	0.044
空間共有コミュニケーション	0.031	高速仮説推論	0.040
高速仮説推論システム	0.031	自然言語	0.037
東京大学工学部	0.030	擬人化エージェント	0.036
マルチモーダル擬人化エージェント	0.030	電子情報工学科	0.035
キーワード抽出法	0.030	北陸先端科学技術大学院	0.035
仮説推論システム	0.029	東京大学生産技術研究所	0.035
感性基盤機能	0.028	北陸先端科学技術大学院大学	0.034
大澤幸生	0.028	マルチモーダルメディア	0.034

語 w のコンテクストワード c のもとでのキーワードとしてのスコア $Score(n,c,w)$ を，次のように与える．

$$Score(n,c,w) = J(n,w) + \alpha J(c,w)$$

$Score(n,c,w)$ が高いほど語 w は人物 n と関連の深いキーワードとなる．なお，α はコンテクストワードの Jaccard 係数の重みであり，コンテクストの影響度合いを調整するものである．$J(n,w)$ が小さいにもかかわらず $J(c,w)$ が大きくなるような語を除くために $J(n,w)$ がある閾値 k_0 以下となる語はキーワードから除外する．また，$|W|$ や $|N \cup W|$ などが閾値 k_1 以下の場合はキーワード除外する．重みや閾値はヒューリスティックに与えている．

提案手法を用いた例として，コンテクストワードをそれぞれ「人工知能」,「大学」としコンテクストワードの Jaccard 係数の重み α を変化させた時の石塚満氏のキーワードを表 2 および 3 に示す．ここでは，α のパラメータとして氏名との Jaccard 係数が閾値 k_0 以上である語 w_i に関して氏名とコンテクストそれぞれの Jaccard 係数の総和の比 $r = \Sigma_i(J(n,w_i)/J(c,w_i))$ を用いた．これにより，α が大きいほどキーワードのスコアに対するコンテクストワードの影響が大きくなる．閾値は，それぞれ $k_0 = 0.001, k_1 = 6$ とした．

まず，抽出されたキーワードを見ると人物名，組織名，研究内容などに関する多様な語が抽出されていることがわかる．また，表 2，3 を比較するとコンテクストワードに応じてキーワードの内容が変化していることがわかる．各表において $\alpha = r$ の時，コンテクストワードを「人工知能」とした表 2 では「知識ベース」や「マルチエージェント」などのキーワードがあわられているのに対して，表 3 では「大学院工学研究科」や「東京大学工学部」などコンテクストワード「大学」に関連したキーワード

表 4 各手法の precision，coverage，context precision (被験者数 10 名)

手法	tf	$tfidf$	提案 (コンテクストなし)	提案 (コンテクストあり)
precision	0.13	0.18	0.60	**0.63**
coverage	0.20	0.24	0.48	**0.56**
context precision	0.05	0.04	0.15	**0.19**

があらわれている．さらに各表の列ごとのキーワードの変化に注目するとコンテクストワードの重みを増やすことにより例えば，表 2 では「人工知能学会」,「自然言語処理」や，表 3 では「工学系研究科」のようなコンテクスを反映した特徴的な対象人物のキーワードが上位に来ていることがわかる．

3. 評 価 実 験

提案手法の有効性を示すために，評価実験を行った．人工知能の研究者 10 人を対象とし，研究者の氏名を検索して得られた検索結果上位 10 件の Web ページ (html のみを対象) に対して tf (Term Frequency)，$tfidf$ (Term Frequency-Inverse Document Frequency)，提案手法 (コンテクスワードなし)，コンテクストワードを「人工知能」とした提案手法[*4]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．いずれも，用語抽出の結果得られた平均 1000 語がキーワード抽出の対象となる．

$tfidf$ の計算では，2004 年度人工知能学会全国大会の参加者 567 名の氏名を検索して得られた計 3981 個の html ファイル (1 人につき最大 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．$tfidf$ 値は正規化を行った．

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，各手法から得られたキーワードを混合した．抽出された各キーワードについて，各被験者に以下の質問を行った．

- 質問 1.「自身の研究活動に関連する語をチェックして下さい.」
- 質問 2.「質問 1 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.」
- 質問 3.「質問 1 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.」

最初の質問は各手法の precision を求めるためのものであり，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 とした．2 番目の質問は各手法の coverage を求めるためのものであり，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 として評価した．最後の質問はコンテクストを反映した precision を求めるためのものであり，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) として評価した．

表 4 には全被験者に対する各手法の precision，coverage および context precision を示す．提案手法は precision，coverage ともに tf および $tfidf$ を大きく上回っている．このことは Web のような多様で不均質な文書集合における，語の共起情報の有効性を示している．すなわち，2 章で述べた仮説 1 の検証と考えることができる．

コンテクストを考慮したキーワードの精度である context precision は，被験者のチェック数が少ないため数値自体はかなり低い．しかし，各手法を比べると提案手法がコンテクストに応じた各人のキーワードを最もよく抽出できていることがわかる．このことは文脈を表すコンテクストワードと語の共起を考慮することの有効性を示しており，仮説 2 の検証と考えることができる．

4. 議 論

4·1 提案手法と $tfidf$ の関連

文書からのキーワード抽出では tf や $tfidf$ が一般的な手法である．以下では，提案手法とこれらの手法の関連を考察する．

語 a が出現する文書集合を A，語 b が出現する文書集合を B とする．例えば，a は提案手法における氏名，b は人物のキーワードである．Web ページ全体の集合 N に対して $|N| \gg |A|$，$|N| \gg |B|$ とする．また，文書集合 A をつなげて，ひとつの文書としたものを d_A とする．2 章で述べた $tfidf$ の適用は A を氏名を検索した結果得られた上位の Web ページ集合，B を複数の氏名を検索して得られた Web ページ集合としたものであり，以下で述べるような Web ページ全体の集合を考慮した時と比べるとかなり限定されたものである．

さて，d_A における語 b の $tfidf$ 値は次の式で計算できる．

$$tfidf(b, d_A) = tf(b, d_A) \times log(\frac{|N|}{|B|} + 1)$$

ここで

- (仮定１)A の各文書中で b は高々1 回しか出現しないと仮定すると d_A 中で語 b の出現頻度 $tf(b, d_A)$ は $|A \cap B|$

*4 パラメータは閾値をそれぞれ $k_0 = 0.001, k_1 = 6$ とした．なお，各人に依存したパラメータを除くために $\alpha = 1$ とした．

```
<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>
......
```

図 3 抽出されたキーワードを利用した FOAF ファイルの例

図 4 研究者検索システム

で表される[*5]．さらに氏名と比べると，キーワードは一般的である．すなわち

- (仮定 2) $|A| \ll |B|$

を仮定すると $|B| \simeq |A \cup B|$ であるので，$tfidf$ 値は 2 つの仮定の下で最終的に次のように近似できる[*6]．

$$tfidf(b, d_A) \simeq |A \cap B| \times log(\frac{|N|}{|A \cup B|}) \qquad (1)$$

一方，提案手法で用いた Jaccard 係数は，

$$J(a,b) = \frac{|A \cap B|}{|A \cup B|} \qquad (2)$$

により語 a と語 b の結びつきの強さを特定するものである．ここで式 (1) と式 (2) を見比べると，$|A \cup B|$ が大きいほど値は小さくなり，$|A \cap B|$ が大きいほど値は大きくなるという点で，両式は同じ意味を持つ．したがって，提案手法は,「キーワード a が出現する文書集合をひとつの文書と考えたとき，仮定 1，仮定 2 のもとでの $tfidf$ に近いものである」と考えられる．同様に，共起頻度 $|A \cap B|$ は,「キーワード a が出現する文書集合をひとつの文書と考えたときの，仮定 1 のもとでの tf 値を表している.」と考えられる．なお，式 (1) では，全文書数 N が必要になり，Web の場合にこの正確な値を出すことは難しいのに対し，式 (2) ではこれを必要としていないことは本手法の特徴のひとつである．χ^2 値 [松尾 02]，対数尤度比 [Dunning 93, 原田 03]，相互情報量 [Church 89] などの共起尺度は，2 つの確率変数間の依存性の程度を測定するため，同様に全文書数の把握が問題となる．

4·2 提案手法の有効性と限界

提案手法では，検索エンジンを用いて氏名を検索する際に，同姓同名が問題となる．同姓同名のために，氏名の検索結果に対象以外の人の Web ページが含まれてしまったり，氏名と語のヒット件数が大きくなってしまうことがある．松尾らは，この問題に対して氏名とともにその人の所属情報を用いている [松尾 05]．佐藤らは人間関係をもとに Web 上での同姓同名の分離を行う手法を提案している [佐藤 04]．本手法では，松尾らと同様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．研究者の場合は，所属情報の取得は難しくないが，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．今後は同姓同名の問題解決の方法を検討する必要がある．

本手法は，研究者や著名人のように Web 上に多くの情報が存在する人や人間関係に関する情報の抽出に特に有効である．例えば，[Dingli 03] では，特定の Web ページおよび属性情報に焦点をあてて情報抽出を行っているのに対して，本手法では個人ページ，Weblog，学会ページ，論文，組織ページ，プロジェクトページなどさまざまな Web ページから多様な属性情報を関連度とともに抽出可能である．表 2 に示すようにキーワードには人物の名前，組織の名前，学会名，研究分野や研究テーマなどさまざまな属性の語が含まれている．人物名や組織名は固有表現[*7]抽出ツール [工藤 02] を使えば判別できるが，その他の属性は固有表現抽出と別に判別を行う必要がある．その際にもキーワー抽出の同様に語の共起情報が有効であると考えられる．例えば，学会に関連する語であれば「会議」,「研究会」,「ワークショップ」などの特徴的な語と共起するはずである．属性の種類の決定あたっては，人物やその関係を表現する FOAF (Fried of a Friend) [Brickley 04] のようなオントロジーが利用できる．例えば，人物名，組織名，研究テーマ，研究プロジェクトなどは `foaf:knows`, `foaf:currentorganization`, `foaf:interest`, `foaf:currentproject` などの属性と対応付けられる [Mori 04]．図 3 はキーワードをもとに作成した研究者の FOAF ファイルの例である．ただし，

*5 なお，仮定 1 は,「A の各文書中での b の出現回数は $|A \cap B|$ と比例している」とおくこともできる．

*6 実際に，仮定 2 は成り立たないことも多いが a と b に関連が強いほど，$|A \cup B|$ は $|B|$ に近づき，関連が弱いほど $|B|$ より大きくなるので，全体の議論には影響しない．

*7 IREX プロジェクト [ire] の固有表現抽出タスクでは，8 種類の表現，組織名，政府組織名，人名，地名，固有物名，日付表現，時間表現，金額表現，割合表現を定義している．

表 5 「松尾豊」と「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

すべての属性が既存のオントロジーで表現されるわけではないので今後はキーワードの属性の自動判別とそれに対応する人物や人間関係を表現するオントロジーの拡張を行う必要がある．また，キーワードの整理のために関連する語をまとめるクラスタリングなども考えている．

提案手法は Web ページにおける語の共起という統計的な情報に基づいているために対象となる情報が Web 上に少ない場合は，高い精度を出すことは難しい．今後，Weblog などの個人コンテンツ作成ツールにより，Web 上に個人に関連した情報が増加すれば，提案手法は適用範囲が広がると考えられる．一方で，Web 上に公開されている情報であるからといって，自由に利用してよいとは限らない．情報を公開している人が想定する範囲を越えて情報を加工，流通させることは問題があるため，プライバシーに十分配慮を行ったうえで情報の抽出を行う必要がある．人物以外では，商品名や企業名に関するキーワード抽出に適用できるだろう．

4·3 提案手法の応用

提案手法は，特定のコンテクストにおける人物のキーワードを抽出するものであるが，コンテクストとして人物の氏名を用いることもできる．この場合，キーワードは 2 者間の関係を表す語となる．例えば，表 2 は松尾豊氏と石塚満氏の関係キーワードである．共通の研究室名や両者をつなぐ人物名が得られている．

コミュニティにおける人の関係の俯瞰を与えたり，人と人をつなぐ関係のパスの発見を効果的に支援するものとして，社会ネットワークが近年注目をあつめている．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 [松尾 05]．この手法を利用して著者らは，研究分野の俯瞰や研究者間の効果的な協働関係の構築のために，Web 上にある研究に関する情報を統合した研究者検索システム POLYPHONET を開発している (図 4)．POLYPHONET は，人のつながりに着目し，Web から抽出した研究者のネットワークを用いて研究者の検索を行うことができる．提案手法により抽出された研究者や研究者間の関係キーワードは，研究者の研究内容や関係性の把握，そしてキーワードを利用した研究者の検索などに利用されている．

5. 関 連 研 究

本研究は，人物に関連した情報を Web から抽出するものである．Web 以外の情報源を対象としたものとしては，論文データベースからの著者の所属抽出 [Han 03] や社内の業務文章からの従業員情報の抽出 [井形 04] がある．これらは，定型的な文書を対象とした情報抽出である．

一方，Web を対象としたものとして，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手法を提案している [山本 00]．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いる．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 [松平 04]．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 [Alani 03]．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．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 [Dingli 03]．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．

これらの従来手法の多くが Web ページの構造に依存したテンプレートやヒューリスティックルール用いたり事前に定義したオントロジーに基づくなどしており，人物名，所属，メールアドレスなどの特定の情報に特化したものになっている．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，事前にテンプレートやルールを与えることは抽出可能な情報を限定してしまうことになる．本手法では，Web ページの構造やドメインに依存せず，氏名のみを入力に用いて，検索エンジンを利用することで人物および人間関係に関する多様な情報抽出を可能にするものである．

検索エンジンにより 2 つの語の間の関係性の強さを求める方法は，最近ではいくつかの研究で用いられている [松尾 05, Mika 04]．これらは Kautz らの Referral Web の研究に基づいている．Referral web [Kautz 97] は，Web ページから人名を収集し人間関係の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ものである．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 [原田 03] がある．これは，分野を表す語に関連する人を Web から抽出するというアプローチであり，人物のキーワー

ドを Web から抽出する本手法とは目的が異なる．

語の共起を用いたキーワード抽出手法は従来，論文や新聞記事などの文書を対象に適用されている [松尾 02, 大澤 99]．提案手法は，Web 上の文書を対象とし，Web に特徴的な検索エンジンを活用したキーワード抽出である．また，コンテクストという概念を導入し Web 上の多様な情報から目的に応じたキーワード抽出を実現している．

6. む す び

本論文では，語の共起情報を用いて Web 上から人物のキーワードを抽出する手法を提案した．また，多様な Web の情報から目的に応じたキーワードを抽出するためにコンテクストの概念を提案し，評価を通して従来手法と比べて有効性を検証した．提案手法の特徴の 1 つは，Web を巨大なコーパスと見なして検索エンジンを用いて語の共起情報を利用している点である．Web 全体を大きなデータベースとみなし情報統合を行うアプローチは，さまざまな情報が電子化され Web からアクセスできるようになったこと，さらに検索エンジンが広く利用可能になったことにより有効性を増している．

Weblog や Wiki などのツールの普及により，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を統合することで高次の情報を抽出し，ユーザの文脈に応じた情報支援を行う方向性は大きな可能性を秘めているのではないだろうか．

謝 辞

本研究について有益なアドバイスをいただいたローザンヌ連邦工科大学 Boi Faltings 氏，国立情報学研究所 武田英明氏に感謝いたします．また，実験にご協力いただいた方々に感謝いたします．

◇ 参 考 文 献 ◇

[相澤 00] 相澤 彰子：語と文書の共起に基づく特徴度の数量的表現, 情報処理学会論文誌, Vol. 41, No. 12, pp. 3332–3343 (2000)

[Alani 03] Alani, H.: Automatic Extraction of Knowledge from Web Documents, in *Proc. of the Workshop of Human Language Technology for the Semantic Web and Web Services, 2nd Interational Semantic Web Conference* (2003)

[Brickley 04] Brickley, D. and Miller, L.: in *FOAF: the 'frined of a friend' vocabulary, http://xmlns. com/foaf/0.1/* (2004)

[Church 89] Church, K. W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, in *Proc. of ACL '89, Association of Computational Linguistics*, pp. 76–83 (1989)

[Davis 04] Davis, I. and Jr., E. V.: in *RELATIONSHIP: A vocabulary for describing relationships between people, http://vocab.org/relationship/* (2004)

[Ding 05] Ding, L., Zhou, L., Finin, T., and Joshi, A.: How the Semantic Web Is Being Used An Analysis of FOAF Documents, in *Proc. of the 38th Ann. Hawaii International Conference System Sciences* (2005)

[Dingli 03] Dingli, A., Ciravegna, F., Guthrie, D., and Wilks, Y.: Mining Web Sites Using Usupervised Adaptive Information Extraction, in *Proc. of the 10th Conference of the European Chapter of the Association for Computational Linguistics* (2003)

[Dunning 93] Dunning, E. T.: Accurate methods for the statics of surprise and coincidence, Vol. 19, No. 1, pp. 61–74 (1993)

[Han 03] Han, H.: Automatic Document Metadata Extraction using Support Vector Machines, in *Proc. of the ACM IEEE Joint Conference on Digital Libraries* (2003)

[原田 03] 原田 昌紀, 佐藤 進也, 風間 一洋: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)

[井形 04] 井形 伸之, 小櫻 文彦, 片山 佳則, 津田 宏：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)

[ire] Information Retrieval and Extraction Exercise, http://www.csl.sony.co.jp/person/sekineIREX/NE/

[Kautz 97] Kautz, H., Selman, B., and Shah, M.: The Hidden Web, *AI Magazine*, Vol. 18, No. 2, pp. 27–36 (1997)

[工藤 02] 工藤 拓, 松本 裕治：チャンキングの段階適用による日本語係り受け解析, Vol. 43, No. 6, pp. 1834–1842 (2002)

[Manning 99] Manning, C. D. and Schutze, H.: *Foundations of Statistical Natural Language Processing*, MIT Press (1999)

[松平 04] 松平 正樹, 上田 俊夫, 大沼 宏行, 渕上 正睦, 森田 幸伯：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)

[松尾 02] 松尾 豊, 石塚 満：語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム, 人工知能学会誌, Vol. 17, No. 3, pp. 213–227 (2002)

[Matsuo 04] Matsuo, Y., Hamasaki, M., Mori, J., Takeda, H., and Hasida, K.: Ontological Consideration on Human Relationship Vocabulary for FOAF, in *Proc. of the 1st Workshop on Friend of a Friend, Social Networking and Semantic Web* (2004)

[松尾 05] 松尾 豊, 友部 博教, 橋田 浩一, 中島 秀之, 石塚 満：Web 上の情報からの人間関係ネットワークの抽出, 人工知能学会誌, Vol. 20, No. 1, pp. 46–56 (2005)

[Mika 04] Mika, P.: Bootstrapping the FOAF-Web: an experiment in social networking network mining, in *Proc. of 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web* (2004)

[Mori 04] Mori, J., Matsuo, Y., Ishizuka, M., and Faltings, B.: Keyword Extraction from the Web for FOAF Metadata, in *Proc. of the 1st Workshop on Friend of a Friend, Social Networking and Semantic Web* (2004)

[大澤 99] 大澤 幸生, ネルス E, 石塚 満：語の共起グラフの分割・統合によるキーワード抽出, 電子情報通信学会, Vol. J82-D-I, No. 2, pp. 391–400 (1999)

[佐藤 04] 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, in *DBWeb* (2004)

[武田 04] 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)

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

[徳永 98] 徳永 健伸：情報検索と言語処理, 東京大学出版会 (1998)

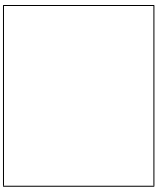
[vca] Representing vCard Objects in RDF/XML, http://www.w3.org/TR/2001/NOTE-vcard-rdf-20010222/

[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)

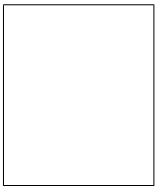
〔担当委員：××○○〕

19YY 年 MM 月 DD 日 受理

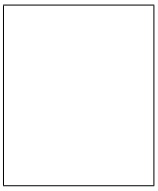
著 者 紹 介



森 純一郎(学 生 会 員)



松尾 豊(正 会 員)



石塚 満(正 会 員)