2E11　Ｗｅｂからの研究者ネットワーク抽出と研究者検索システム

○松尾　豊（産総研），浅田洋平，森純一郎（東大），
石黒　周（研究開発型ＮＰＯ振興機構），松原　仁（はこだて未来大），橋田浩一（産総研）

1　はじめに

最近では，研究に関するさまざまな情報がWebから手に入る．例えば，研究者個人の研究に関する内容の紹介や発表文献、学会のプログラム，プロジェクトや研究グループのホームページ，採択された助成金の情報など，多様な情報がWeb上に存在する．我々は、研究者に関する情報をWeb上から集め、その関係を抽出する手法を研究している．これまで人工知能学会の研究者の関係を抽出し図示するシステムを2003年度と2004年度の人工知能学会全国大会において運用した．学生や若手研究者，他分野の研究者が、当学会内の研究者の関係を把握したり、研究分野を俯瞰する用途に用い、好評を博している。

一方、近年では、産学官連携の重要性がますます高まっている。NPO型分散研究システムでは、NPOが中核となって，自律分散的な研究者がネットワークされ研究ゴールをめざすという形での研究システムが提案されている。こうした仕組みの構築にあたって，我々が研究を進めている研究者ネットワークの抽出技術が何らかの貢献ができると考えている．自分に馴染みの少ない研究分野の研究者ネットワークを自動的に抽出し、どういった研究者がどのようなグループを構成しているか、どういう研究テーマが行われているといった全体像を俯瞰することは，研究者や事業者などさまざまな主体の交流に役立つのではないだろうか．また，実際に研究者ネットワークが変化していく様子を捉えることができれば，活動の評価や方向性の決定にも使えるのではないだろうか．

これまで，論文DBの共著や引用関係を用いて研究者の関係を分析する研究は多く行われてきた．しかし，Web上には，発表文献やプロジェクトの情報を含んだ，より多様な情報が存在し，非常に新しい情報も含まれる．例えば，研究の開始からその成果が論文となって公表されるには1年以上かかのが普通だが，研究を始めた時点でその目的や内容をWeb上で紹介することも珍しいことではない．我々は，Web上にある情報の多様性やその鮮度を重視し，特にWebを対象として技術開発を進めている．学会におけるコミュニケーション支援や，研究者の検索，効果的な協働研究の促進が大きな目的である．

以下，Webからの研究者ネットワークの抽出技術 [松尾 04a, Matsuo 04b]、およびそれを用いた研究者検索システムについて述べる。

2　研究者ネットワークの自動抽出

2.1　関係の強さの抽出

ここでは、ネットワークの抽出法を人工知能学会の研究者を例にとって説明する。

まず、ネットワークを構成するのは、2004年度の人工知能学会の全国大会 (JSAI2004) の著者・共著者とし，ネットワークのノードとする．ネットワークに含める研究者は、あらかじめ目的とする研究コミュニティの研究者リストを何らかの方法で入手しておけばよい。なお，本手法では、個人に関する情報として用いるのは，氏名と所属だけである．

次に、ノード間にエッジを付与する。基本的なアルゴリズムは非常にシンプルである．例えば，「松尾豊」と「石塚満」の関係を調べるときには，検索エンジンに “松尾豊 AND 石塚満” と入力する．「松尾豊 AND 石塚満」の場合には，156件のヒットがあるのに対し[1]，「松尾豊 AND 溝口理一郎」の場合には7件のヒットしかない．「石塚満」単独では1120件のヒット件数，「溝口理一郎」単独では1130件のヒット件数であり，ほぼ同数であるから，「松尾豊」とANDをとったときの件数の違いは，氏名の共起関係の強さの違いを表していると考えることができる．

氏名が共起するページというのは，研究室のメンバーのページ，業績リストのページ，論文データベース，学会や研究会のプログラム，大学内の教官メンバーリストなどさまざまである．そして，このようなページが多くあるほど，両

[1] 2004年1月8日時点でのGoogleによる検索結果．以下の例でも同様．Googleでは姓と名の間をつめて正確な氏名の検索が可能である．

者が何らかの社会的関係にあり，またその関係が強い可能性が高いというヒューリスティックを本研究では用いている。

本システムでは、共起の強さを測る指標として、つぎの Simpson 係数（もしくは Overlap 係数）を用いる。

$$R(X,Y) = \begin{cases} \dfrac{|X \cap Y|}{min(|X|,|Y|)} & \text{if } |X| > k \text{ and } |Y| > k, \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$R(X,Y)$ は,「X」と「Y」の関係の強さを表す関数であり，k は閾値である．JSAI2004 の場合，$k = 30$ とした．つまり単独でのヒット件数が 30 件以下の人はエッジが張られない．

また，同姓同名の問題に対処するために、氏名とともに所属もクエリとして用いた。例えば,「松尾豊」の場合には、“松尾豊 産業技術総合研究所” というクエリを用い検索する．なお，複数の所属機関にまたがっている場合や所属が変わった場合は，それらを OR でつなげたものを用いる．また，東大と東京大学など，代表的な機関の略称や別名については，同義語辞書を作り，同義語拡張を行った上で検索を行う．

2.2 関係の種類の抽出

次に，検索にヒットしたページから関係の種類を判別する。研究者の関係の種類として，本システムでは次のようなクラスを定めた．

共著関係 共著の論文がある関係．

同研究室関係 同じ研究室や研究所のメンバーなど所属が同じである (あった) 関係．

同プロジェクト関係 同じプロジェクトや委員会など，組織をまたがる同グループに所属している (いた) 関係．

同発表関係 同じ研究会で発表する (した) 関係．

ひとつのエッジは複数のラベルを持つことができる．

このような関係を抽出するために，まず検索エンジンに「X and Y」をクエリとして入力し、上位 5 ページを取得する。次に、それぞれのページから属性の値を抽出する．ここでいう属性とは、例えば、X と Y が同行内で共起したか、X および Y の出現回数、タイトルや最初の 5 行に別に定義した語群に含まれる語が出現するかなどである．この属性を用い，判別ルールによって共著や同研究室などどのクラスにあたる関係かを判断する．この判別ルールは，あらかじめ人手で付与した訓練例を用い，C4.5 を用いて生成する．

2.3 研究者キーワードの抽出

研究者間のつながりの強さやその関係の種類だけでなく，各研究者がどのような研究をしているかなどを表すキーワードがあれば，その研究者を理解するのに役立つ．また，2 人の研究者間の関係のキーワードがあれば，例えば，この 2 人は同じ研究室の出身であるとか，同じ研究者とよく研究をしているなどという情報が分かって便利である．ここでは，このような研究者に関するキーワードを研究者キーワードと呼ぶことにする．

研究者キーワードを求めるには，まず氏名（および所属）を検索エンジンにクエリーとして入力し，検索結果の上位 10 件を取得する．それらのページに含まれる語を専門用語抽出ツール Termex を用いて抽出する．こうして抽出した語が，研究者のキーワード候補となる．キーワードは，コミュニティの文脈に合致していた方が望ましい．例えば人工知能学会の研究者なら「人工知能」，ロボット学会なら「ロボット」のように，コミュニティの文脈を表す語をここではコンテクストワードと呼ぶことにする．キーワード候補の中から選んだ語 a に対し，語 a と研究者の氏名，および語 a とコンテクストワードの関連度を検索エンジンのヒット件数を用いて測り，両方の関連度が強い語 a を研究者キーワードとして抽出する．また，コンテクストワードとして，他の研究者の氏名をいれることで，2 人の研究者に関連の深いキーワードを抽出することができる [Mori 04].

2.4 研究カテゴリの抽出

目的とする研究者コミュニティにおいて，研究者の研究分野内における研究カテゴリは，それほど明確に分かれていない場合が多い．学会には通常，研究カテゴリ表などの分類があるが，同じ研究者でも徐々に研究テーマがシフトして

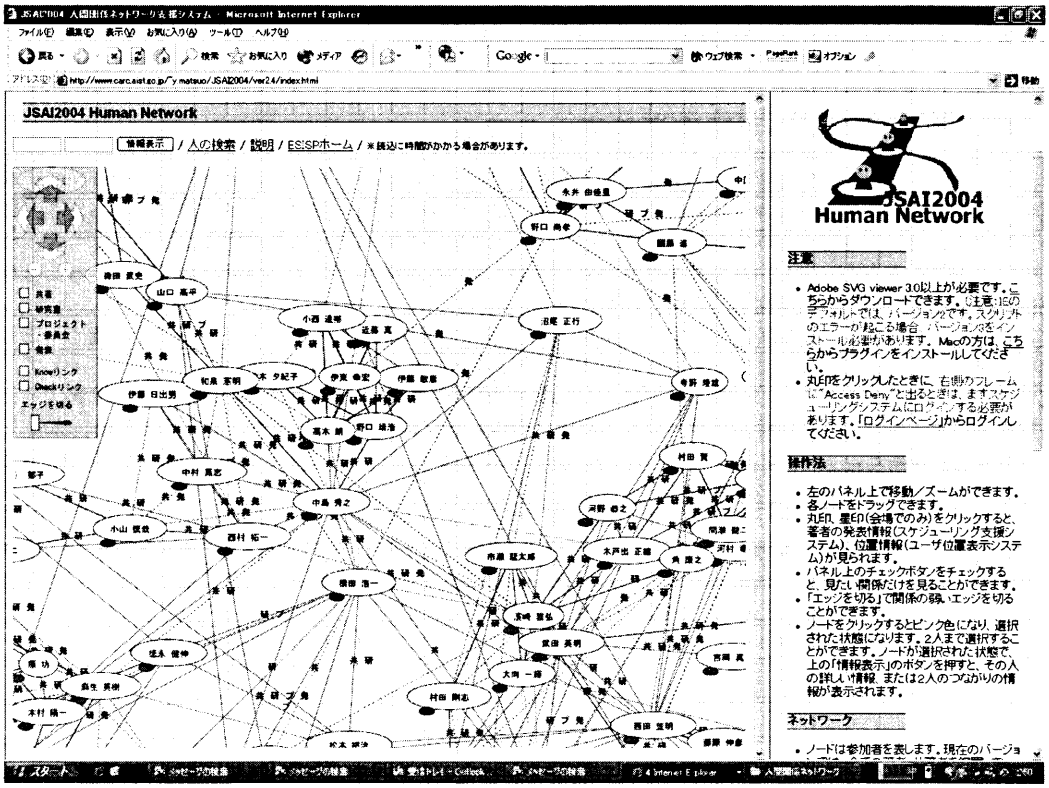

図 1: JSAI2004 で表示した人間関係ネットワーク

いく場合もあれば，複合的な課題を研究している場合もある．

そこで，Web 上の情報を用いて，研究者の分類も自動的に行うことを考える．まず，研究で用いられることの多い一般的なキーワードを用意する．(分類キーワードとよぶことにする．) 分類キーワードは，学会の論文のタイトルやその内容に含まれる頻出語などを用い，論文のテキストがあれば自動的に得ることができる．そして，この分類キーワードと研究者の氏名の共起の強さを，検索エンジンのヒット件数により取得する．分類キーワードと研究者の集合に対して，共起の強さを調べることによって，共起行列を得ることができる．この共起行列に対して，co-clustering とよばれる処理を行うことで，自動的に研究者のグループ，分類キーワードのグループができることになる [Asada 04].

2.5 JSAI2004 におけるシステム

JSAI2004 では，研究者のネットワークを，会場内に設置された KIOSK 端末および Web 上で表示するサービスを行った．表示したネットワークを図 1 に示す．ノード数 275，エッジ数 583[2] のネットワークである．JSAI2004 の著者、共著者の計 567 名から，単独でのヒット件数が閾値に満たない人，他と関係の弱い人を除いた 275 名から構成されるネットワークである．

ネットワークは，SVG[3] で出力され，SVG viewer により閲覧することができる．Javascript が埋め込まれているので，ノードをドラッグしてつながり具合を確かめることができる．各ノードには丸印のアイコンがあり，スケジューリング支援システムと連携している．エッジは，Simpson 係数 $R(X,Y)$ が閾値を越えるノードペア X，Y に対して実線で表示している．破線のエッジはそれよりも閾値が低いもの，赤線のエッジは共起件数自体が大きいものである．エッジラベ

[2] 破線エッジ 171，赤エッジ 174.
[3] SVG は，W3C によって作成された規格であり，ベクトル表現による XML 形式のグラフィック記述言語である．

ルとして，“共”（共著）、“研”（研究室）、“プ”（プロジェクト）、“発”（発表）が付与されている．初期配置では，エッジの長さが $R(X,Y)$（の逆数）をできるだけ反映するような配置となっている．

3 人のつながりを用いた研究者検索システム

我々は，他分野の研究者や研究者以外の人が，自分の要望に適した研究者をうまく検索するための研究者検索システム（仮称：Polyphonet，ポリフォネット）を構築中である．現在，他の研究分野の人と共同研究を行ったり，研究の話を聞いたりするために，自分の知りあいに連絡をとったり，知り合いを通じて適切な研究者を紹介してもらうなどの形が多いのではないだろうか．もし，自分の知り合いと，目的とする研究者がどのような関係かを理解することができれば，連絡も取りやすいし，共同研究もしやすくなるだろう．

本検索システムは，次のような点を特徴としている．まず，氏名や所属，研究キーワードや研究分野をキーとして，研究者の検索を行うことができる．研究キーワードや研究分野はWebから自動的に抽出したものである．そして，検索した研究者がどういった研究者とつながりが深いのか，共著や同研究室関係にある研究者は誰なのかを閲覧することができる．順次，研究者をたどっていくことで，コミュニティ全体の研究者の関係を概観することができる．

また，つながり検索という機能を用いると，ある研究者から別の研究者へのパスを検索することができる．例えば，自分からある研究者へどのようなパスで到達できるのかといったことを調べることができる．

本検索システムで検索の対象となるのは，人工知能分野やロボット分野など，あらかじめリストを与えてWeb上から情報を抽出しておいた研究者である．しかし，場合によっては探したい研究者や自分自身がデータベースに含まれていないこともあり得る．そのため，このシステムでは，自分が関係を見たい研究者を新しく登録することができる．Webから情報を抽出し統合する処理のために，10分〜20分程度の時間はかかるが，登録した研究者が新たにデータベースに追加される．現在は，人工知能やロボットの分野を対象としてシステムを構築しているが，今後，さまざまな研究分野に適用することができると考えられる．

4 おわりに

本稿では，研究者の関係とそれに付随するさまざまな情報をWebから取り出す手法を簡単に紹介した．今後，研究に関するますます多くの情報がWeb上に置かれるようになると考えられるが，こういった情報をうまく統合し処理することにより，研究者のネットワークや研究に関連するより多くの情報を精度良く取り出すことが我々の目標である．この研究が，NPO型分散研究システムなどの効果的な研究推進の仕組みづくりに貢献できるよう，研究開発を進めていきたいと考えている．

参考文献

[Asada 04] Asada, Y., Matsuo, Y., and Ishizuka, M.: A method to automatically find foaf:Group based on the cooccurrence of people with keywords in the Web, in *Proc. 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web*, pp. 34–37 (2004)

[松尾 04a] 松尾 豊, 友部 博教, 橋田 浩一, 中島 秀之, 石塚 満：イベント空間支援における人間関係ネットワーク抽出技術の活用, 人工知能学会全国大会, No. 3C1-04 (2004)

[Matsuo 04b] Matsuo, Y., Tomobe, H., Hasida, K., and Ishizuka, M.: Finding Social Network for Trust Calculation, in *Proc. 16th European Conference on Artificial Intelligence (ECAI2004)*, pp. 510–514 (2004)

[Mori 04] Mori, J., Matsuo, Y., Ishizuka, M., and Faltings, B.: Keyword Extraction from the Web for FOAF Metadata, in *Proc. 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web*, pp. 1–8 (2004)