1A3-05

研究者ネットワーク抽出検索システム

Human Network Search Engine

松尾 豊[*1] Yutaka Matsuo
石田 啓介[*1] Keisuke Ishida
森 純一郎[*2] Junichi-Lo Mori
友部 博教[*3] Hironori Tomobe
石黒 周[*4] Shu Ishiguro
松原 仁[*5*1] Hitoshi Matsubara

[*1]産業技術総合研究所 National Institute of AIST
[*2]東京大学 The University of Tokyo
[*3]名古屋大学 Nagoya University
[*4]研究開発型 NPO 振興機構 R&D NPO Institute
[*5]はこだて未来大学 Future University, Hakodate

Recently, a vast amount of information related to research activities is available on the Web. We developed a social network extraction system, that is based on Web mining technology and can find relationships among researchers such as coauthorship and same laboratory. In this paper, we overview our approach and introduce a researcher retrieval system based on the extracted social network. It is very important to find appropriate researchers to collaborate in an interdisciplinary research field. Also, the cooperation with local communities and industrial communities is important for researchers. Our goal is to support and promote the efficient collaboration so that research activities have greater impacts on our society.

1. はじめに

医療、ロボット、ライフサイエンス、防災、環境などさまざまな分野で、研究者の連携が必要とされている。ロボットではデバイス、ソフトウェア、言語処理、通信などの専門家が協調する必要があし、防災分野では、地震、建築、土木、都市計画などの研究者と、行政や地域がうまく協調する必要がある。しかしながら、異なる研究分野の連携は、現状では研究者の個人的な知り合い関係や紹介によって行われるのが常であり、システマティックに効率的な連携が生まれているとは言いがたい。日本は米国や EU と比較し、人口当たりの研究者人数は多いものの、研究者総数では少ない。また、政府系の研究費も 3 倍から 4 倍の開きがある。絶対的に少ない研究資源を使って、先端的な科学技術分野で国際競争力を上げていくためには、効率的な研究体制、特に複合領域における効果的な連携体制を、戦略的に構築していく必要があると考えられる。

昨今では技術分野の専門化、多様化が進み、また技術進歩も目まぐるしいため、各研究者が効果的な連携体制を作るためには、情報技術による支援が不可欠である。論文 DB や研究者の登録情報などを用いて研究者を検索するシステムは存在するが、研究が論文として公刊されるのは実際の活動より数年遅れであるし、新しい技術分野は既存のカテゴリに収まらない形で現れてくる。したがって、研究者および産業のダイナミックな変化に対応でき、詳細な専門分野を特定した上で共同研究者を探すことができ、しかも研究者の当専門分野内における立場や役割なども考慮できるといった要件を供えた情報支援が必要である。

我々はこれまで、多種大量の情報が存在する WWW（Web）に着目し、そこから情報を抽出する研究を行ってきた。今後もビジネス、コミュニケーション、個人の情報発信など、社会や個人の活動に関わるますます多くの情報が Web 情報空間に存在するようになるだろう。研究者にとっても、研究活動の紹介や研究業績を Web 上で公開する機会も増えており、研究に関する情報が最も早く公開されたのが Web であったということも多くなっている。全世界に数十億の Web ページがあるなかで、単なる情報検索ではなく、どのように情報を集約していくかが今後、最も大きな課題である。本研究では、Web 上にある鮮度の高い情報と、その情報を集約する高い技術を背景に、複合領域において研究者が効果的に研究成果を挙げていくにはどうすればよいか、また分野全体としてどのような戦略を持って研究を推進していけばよいかという知見を得、研究者の効果的な共同研究のための情報支援を行うことを目標としている。

我々は、昨年、一昨年と人工知能学会全国大会において、イベント空間情報支援プロジェクトの一環として、研究者ネットワークの可視化、位置情報やスケジューリング情報と連携した情報提示を行ってきた．今年度は，これを結実させた形で，研究者ネットワーク抽出検索システムとして運用を行うことを予定している．

以下では，Web からの研究者ネットワークの抽出技術 [2, 7]、およびそれを用いた研究者ネットワーク抽出検索システムについて述べる。

連絡先: 氏名，所属，住所，電話番号，Fax 番号，電子メールアドレスなど

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簡単に説明する。

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．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．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、Simpson 係

数（もしくは Overlap 係数）を用いる。

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

$R(X,Y)$ 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．

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る。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．このような Web 上での同姓同名の問題は，Web 上で実体の情報を検索するという重要な問題のひとつのインスタンスであり [5]，本質的な解決は今後の重要な課題のひとつである．

検索エンジンへの負荷を減らすことは，より大規模な研究者ネットワークを抽出する上で重要な課題である．n 人のノードに対して，検索クエリーが $O(n^2)$ 回必要になるが，これを $O(n)$ に落とす手法を開発している．これにより，数千人から数万人の規模のネットワーク抽出が可能である [8].

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，機械学習により生成する．

関係性の判別は，テキスト分類問題の一種と考えることができる．テキスト分類に関しては，近年では SVM を用いる方法やブートストラップ的な手法を用いる方法等，さまざまな手法が提案されている．

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 [3].

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 [1].

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 2.5 に示す．ノード数 275，エッジ数 583[*1]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[*2]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．

*1 破線エッジ 171，赤エッジ 174.
*2 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．

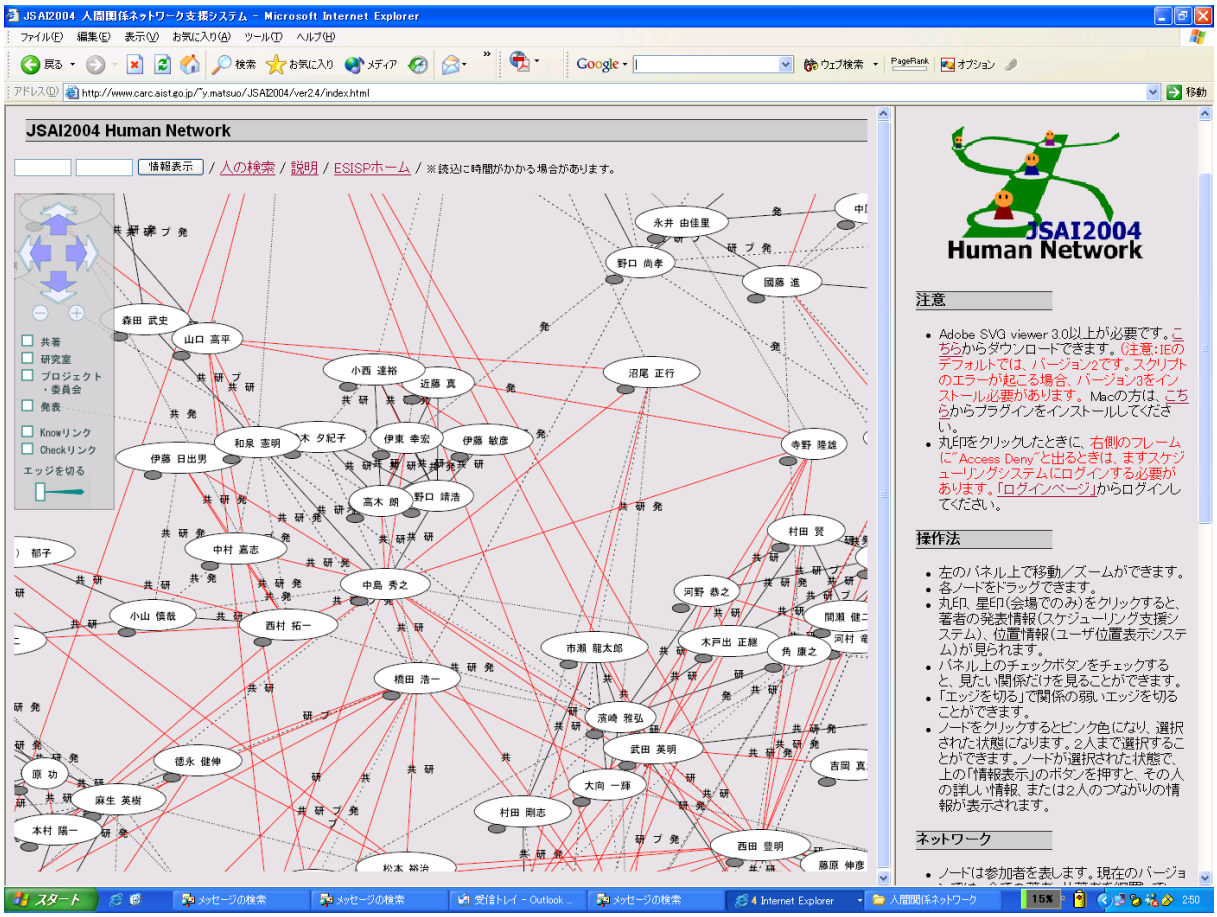

図 1: JSAI2004 で表示した人間関係ネットワーク

3. 研究者ネットワーク検索エンジン Polyphonet

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研究者ネットワークを使ってうまく検索するためのシステム Polyphonet（ポリフォネット）*3を開発している (図 2)．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は、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．

*3 polyphony（多声音楽）+network の造語．

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，30 分〜1 時間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. 今後の課題と方向性

Polyphonet は，イベント空間情報支援における過去 2 年間のシステム構築や，それに伴って必要な技術の研究に基づいて構築されている．今後は，抽出の精度をさらに上げていくことを追求していきたい．その中で，次のような点が課題になる．

- 同姓同名問題など，また表記ゆれの問題など，現実の実体の情報をいかに得るかという課題
- 検索エンジンを使って，Web 全体の情報を効率的に得るための方法．特に，機械学習，キーワードスパイス [4] などの手法をうまく組み合わせて用いる必要がある．
- 構文解析やパターンの処理によって，テキストの内容を詳

図 2: 人工知能版 研究者ネットワーク抽出検索システム

細に分析することにより，情報の確実性を確かめる課題．

Web は，情報源として潜在的に大きな可能性を秘めており，さまざまな研究が行われている [6]．研究者ネットワークに限らず，さまざまなモノに関する情報を Web から抽出し，特にネットワーク的な視点からその性質を明らかにするという方向で研究を進めていきたいと考えている．

5. 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説明し、我々の研究の概要を紹介した。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し、それによって、ユビキタス情報環境における研究者の交流の支援を行っていきたいと考えている．

参考文献

[1] Yohei Asada, Yutaka Matsuo, and Mitsuru Ishizuka. 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.

[2] Yutaka Matsuo, Hironori Tomobe, Koiti Hasida, and Mitsuru Ishizuka. Finding social network for trust calculation. In *Proc. 16th European Conference on Artificial Intelligence (ECAI2004)*, pp. 510–514, 2004.

[3] Junichiro Mori, Yutaka Matsuo, Mitsuru Ishizuka, and Boi Faltings. 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.

[4] Satoshi Oyama, Takashi Kokubo, and Toru Ishida. Domain-specific web search with keyword spices. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, Vol. 16, No. 1, pp. 17–27, 2004.

[5] 佐藤進也, 風間一洋, 福田健介, 村上健一郎. 実世界指向 web マイニングの提案とその同姓同名人物分離問題への適用. 日本データベース学会 Letters, Vol. 3, No. 4, pp. 21–24, 2005.

[6] 藤井敦. 百科事典としてのWWW. 人工知能学会誌, Vol. 19, No. 3, pp. 296–301, 2004.

[7] 松尾豊, 友部博教, 橋田浩一, 石塚満. Web 上の情報からの人間関係ネットワークの抽出. 人工知能学会論文誌, Vol. 20, No. 1E, pp. 46–56, 2005.

[8] 浅田洋平, 松尾豊, 石塚満. Web からの研究者ネットワーク抽出の大規模化. 2005. 投稿中.