3C1-04

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

Utilization of Social Network Extraction for Event Space Information Support n

松尾 豊[*1] 友部 博教[*2] 橋田 浩一[*2] 石塚 満[*3]
Yutaka Matsuo Hironori Tomobe Kôiti Hasida Mitsuru Ishizuka

[*1]産業技術総合研究所 National Institute of Advanced Industrial Science and Technology
[*2]名古屋大学 Nagoya University
[*3]東京大学 University of Tokyo

Social relation plays an important role in a real community, thus it is important for an event space information support. This paper introduces a new approach to automatically obtain a social network of a community from the Web: Nodes are given beforehand. Edges are added consulting to a Web search engine. If two names co-occurs in a lot of Web documents, we assume these two have a strong relation. Moreover, by analyzing the retrieved documents, edge labels are assigned to edges to represent classes of relations such as co-author, same laboratory, same project, or same conference. We operated our system at JSAI2003. Various evaluations are made to show the effectiveness of our approach.

1. はじめに

イベントには，多くの人が集まる．交流を目的としたイベントでは，人そのものが目的の対象であるし，明示的に交流を目的としていないイベントであっても，人と人との交流が重要である場合が多い．JSAI2003 のような学会でも，発表や聴講が主な目的であっても，話をして意見や情報を交換したり，食事に行ったりすることで，結果的に参加者の交流が深まる．このような人の交流を促進するための高度な情報支援システムを作ろうとすると，システムが何らかの形で人と人との関係を把握しておく必要があるだろう．また，センサやアクチュエータが環境中に多数配置された，ユビキタス環境における情報支援を考える際にも，この「人」という要素は重要である．例えば，ユーザの文脈を把握するにあたって,「ユーザがいま誰といるのか」は非常に重要である．恋人といるのか，家族といるのか，友人といるのか，同僚といるのか，先生といるのかは，ユーザの可能な行動や望ましい行為，欲しい情報に大きく関わってくる．

日本では，人間関係を利用して何かお願いする，コネを使うということはあまり好ましいものではないという印象が強いが，実際、共同研究を行う、査読をお願いする、学会を運営するなど、仕事・研究の場面において，人間関係が少なからず活用されることも多い．networking という言葉で表されるように，いかにネットワークを作って自分の活動を効率的に行う環境を整えていくかは，個々の研究者にとって，またひとつの学問分野全体においても重要な視点である．

我々は，人の関係が重要だと考え，人間関係に着目した情報支援を目指している．人間関係という語には，オフィシャルなものからプライベートなものまで幅広いニュアンスが含まれるが，我々が対象とするのは研究者の協働関係やコミュニティにおける友人関係など，情報支援に役立ち，かつプライバシーにできるだけ踏み込まないものである．本稿では，人間関係ネットワーク，すなわち人工知能学会を対象として研究者間の協働関係を Web から抽出する手法，および JSAI2003 での運用とその後の評価について述べる．

連絡先: 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，〒 135-0064 東京都江東区青海 2-41，03-3599-8327，y.matsuo@carc.aist.go.jp

2. 人間関係ネットワークの抽出

人間関係ネットワークを構成するメンバーはあらかじめ決められているとする．例えば，JSAI2003 などの学会の参加者[*1]の氏名は，開催に先だって公開されている．JSAI2003 におけるネットワークの場合には，1999 年から 2003 年まで 5 年間の人工知能学会全国大会における著者および共著者をノードとした．我々が個人に関する情報として事前に用意するのは，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処理を行う．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　石塚満”

と入力する．(両者は AND の関係である.)「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[*2],「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．すなわち,「松尾豊」と「石塚満」の方が,「松尾豊」と「溝口理一郎」よりも同一ページに出現する傾向が強い．したがって、関係が強いであろうことが推測される．実際，この例では，石塚満氏は松尾豊氏の学生時代の指導教官である．なお，本論文では，同一の Web ページに氏名が同時に現れることを，氏名が共起する，ということにする．

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，というのが本研究の仮説である．

*1 厳密には発表論文の著者と共著者で，聴講のみの参加者は含まない．

*2 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．

2.1 共起の強さを正確に知る

共起の強さを測るために，共起頻度以外にもさまざまな指標がある．集合の類似度，重なり具合を表す指標として，さまざまなものが提案されている [Manning 02]．ここでは，氏名「X」と氏名「Y」の単独でのヒット件数をそれぞれ $|X|$，$|Y|$，AND をとったとき，OR をとったときのヒット件数をそれぞれ $|X \cap Y|$，$|X \cup Y|$、Web ページ全体の数を N とする．共起頻度：$F(X,Y) = |X \cap Y|$，相互情報量：$log\frac{N|X \cap Y|}{|X||Y|}$，ダイス係数：$\frac{2|X \cap Y|}{|X| + |Y|}$，Jaccard 係数：$\frac{|X \cap Y|}{|X \cup Y|}$，Simpson 係数：$\frac{|X \cap Y|}{\min(|X|, |Y|)}$，コサイン：$\frac{|X \cap Y|}{\sqrt{|X||Y|}}$ などである．

共起頻度は，単独でのヒット件数が多い人ほど有利という問題がある．一方，他の係数は逆の欠点がある．仮に $|X|$ と $|Y|$ の差が大きい場合を考えよう．例えば，$|X| = 1000$，$|Y| = 30$，$|X \cap Y| = 30$ とすると，Jaccard 係数は 30/1000 と小さな値になる．$|Y|$ から見ると，すべてのページで $|X|$ と共起しているにも関わらず，値が小さい．例えば，研究室の学生と先生では，先生のヒット件数が多いためにエッジが張られないことになってしまう．

ただし，Simpson 係数は，分母に関して min をとっているので，この欠点がない．この係数は，ヒット件数の小さい方から見た距離感を表しており，例えば，研究室の学生と先生の場合にも，学生から見て先生と共起する割合が高ければエッジが張られることになるので，先生がたくさんのエッジを集めることになる．これは，研究室における協働関係に対して，我々の持っている印象と一致する．

しかし，Simpson 係数にも，単独でのヒット数が非常に少ない人には特に高い値が出やすいという欠点がある．例えば，$|X| = 1$，$|Y| = 100$，$|X \cap Y| = 1$ の場合，Simpson 係数は 1 と最大値になる．これは，$|X|$ のサンプル数の少なさに起因して値の推測が粗すぎることが原因である．この欠点を解消するために，我々は次のような閾値つき Simpson 係数を用いることにした．

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

$R(X,Y)$ 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．JSAI2003 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．他にも，統計的な信頼度を推定する、m-estimate 法を用いるなどの方法が考えられるが，ここでは簡単のため，閾値による足切りを行う上式を採用した．

また，単独でのヒット件数，例えば「松尾豊」のヒット件数を得る際，単純に “松尾豊” をクエリとして検索を行うと，容易に想像がつくように同姓同名の松尾豊氏が多数ヒットする．例えば,「松尾豊」の場合，ヒット件数 903 件に対し，本論文の著者である松尾豊氏に関するページは 256 件である．そこで，氏名とともに所属情報を用い，より正確なヒット件数を得るという工夫を行う．例えば,「松尾豊」単独のヒット件数を得るために，“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．JSAI2003 の場合には，学会のプログラムに記載されている所属情報を用いた．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 エッジラベルの抽出

表 1: 「X AND Y」でヒットした Web ページから抽出する属性

属性名	説明	値
NumCo	二人の氏名の共起回数	zero, one, more_than_one
Rel	Simpson 係数が閾値以上か	yes, no
FreqX	X の出現回数	zero, one, more_than_one
FreqY	Y の出現回数	zero, one, more_than_one
GroTitle	タイトルに語群が出現するか	yes, no
GroFFive	最初の 5 行に語群が出現するか	yes, no

※ GroTitle と GroFFive 属性は，語群 A-F に対してそれぞれ定義されるので，両方で 12 個の属性となる．

次に，氏名が共起したページ，つまり検索にヒットしたページの特徴を用いて，関係の種類を判別する手法について述べる．

社会的関係の種類として，本論文では研究分野に特有の次のようなクラスを定める．これらが，エッジのラベルの種類となる．

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．今回は，研究者を対象としているのでこのような関係を定義したが，一般的には対象とする領域ごとに定義する必要がある．

さて，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[*3]．次に、それぞれのページから表 1 にある属性の値を抽出する．GroTitle 属性，GroFFive 属性は，そのページが何に関するページであるかを判断するためのものであり，別に定義した語群（表 2）を用い，語群 A がタイトルに出現するかどうか（GroTitle(A) 属性），語群 B が最初の 5 行に出現するか（GroTitle(B) 属性）などを表す．各語群はあらかじめ正解クラスの付与されたページを用い，各クラスごとに TF·IDF 値の上位語を語群としている．例えば，語群 A は論文リスト・業績リストのページであるか，また語群 B は研究室のページであるか判断するために利用できる．

例を用いて説明すると,「友部博教 AND 石塚満」で検索したあるページ[*4]から 1 の属性を抽出すると，

(more_than_one, yes, yes, more_than_one, more_than_one,
no, no, no, no, no, no,
yes, no, no, no, yes, no)

となる．そして，この属性から共著・研究室・プロジェクト・発表という 4 つのクラスに属するかどうか，このページの場合には (Yes, No, No, Yes) を得るという問題になる．さらに検索にヒットした他のページからも関係を求め、最終的に 2 人

*3 Google の上位候補は PageRank が高い Authority ページであり，またなるべく重複が避けられるように工夫されているので，そのまま上位から優先的に用いる．

*4 http://www-kasm.nii.ac.jp/jsai2003/programs/person-182.html

表 3: 獲得した全判別ルール

クラス	判別ルール
共著	NumCo = more_than_one
研究室	(NumCo = more_than_one & GroFFive(F) = no) or (Rel = yes & GroTitle(E) = no & GroFFive(C) = no) or (GroTitle(A) = yes & GroFFive(C) = no & GroFFive(F) = yes) or (GroTitle(E) = no & GorFFive(B) = yes & GroFFive(C) = no) or (GroTitle(E) = no & GroFFive(B) = yes & GroFFive(E) = yes) or (GroTitle(E) = no & GroFFive(B) = yes & GroFFive(F) = yes)
プロジェクト	(FreqX = one & GroTitle(B) = yes) or (GroTitle(C) = yes) or (GroTitle(C) = no & GroFFive(C) = yes & GroFFIve(D) = no & GroFFive(E) = no) or (Rel = no & FreqX = one & GroTitle(B) = yes)
発表	(FreqY = more_than_one & GroTitle(D) = yes) or (GroTitle(F) = yes & GroFFive(D) = yes) or (NumCo = zero & GroTitle(F) = no & GroFFive(B) = no & GroFFive(E) = no & GroFFIve(F) = yes) or (NumCO = zero & GroTitle(C) = no & GroFFive(D) = yes & GroFFIve(E) = no) or (NumCO = zero & GroTitle(C) = no & GroTTl(D) = no & GroFFive(D) = yes)

表 2: 語群

語群	語
A	出版, 論文, 発表, 活動, テーマ, 賞, 著者
B	メンバー, 研究室, 研究所, 研究機関, チーム
C	プロジェクト, 委員会
D	ワークショップ, 会議, セミナー, ミーティング, スポンサー, シンポジウム
E	学会, 団体, プログラム, 国立, ジャーナル, セッション
F	教授, 専攻, 大学院生, 講義

表 4: ラベルのエラー率，適合率と再現率

クラス	エラー率*	適合率	再現率
共著	4.1%	91.8% (90/98)	97.8% (90/92)
研究室	25.7%	70.9% (73/103)	86.9% (73/84)
プロジェクト	5.8%	74.4% (67/90)	91.8% (67/73)
発表	11.2%	89.7% (87/97)	67.4% (87/129)

の関係のエッジラベルとして (Yes, Yes, No, Yes) つまり、共著かつ研究室かつ発表関係であると求めたい．

したがって，ページの属性から自動的にクラスを判別できればよい．これは，属性からクラスを予測するルールを学習する問題となる．本研究では，C4.5[Quinlan 93] を用いて判別ルールを生成する．ランダムに抽出した 275 ページを人手で正解クラスを付与し，これを訓練例として用いた．獲得したルールを 3 に示す．

共著関係のルールでは，氏名が同行内に出現することが 2 回以上あれば（NumCo=more_than_one）共著と判断する，という非常に簡単なものである．研究室関係を判断するルールでは，例えば 1 つ目のルールは，名前が 2 回以上共起している（NumCo=more_than_one）のに学科や講義のページではなければ（GroFFive(F)=no）研究室関係である，というルールである．

表 4 に，275 の訓練例を 5 群に分け，クロスバリデーションを行った平均エラー率を示す．また，実際に得られたラベルの適合率，再現率を知るために，ルールを生成する際に用いた 275 の訓練例とは別にランダムに 200 個のエッジを選び，そのラベルを人手で判定し，適合率，再現率を求めている．およそ 7 割から 9 割の適合率，再現率となっている．

3. JSAI2003 におけるシステムと評価

3.1 人間関係ネットワークの表示

JSAI2003 では，人間関係ネットワークを，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．

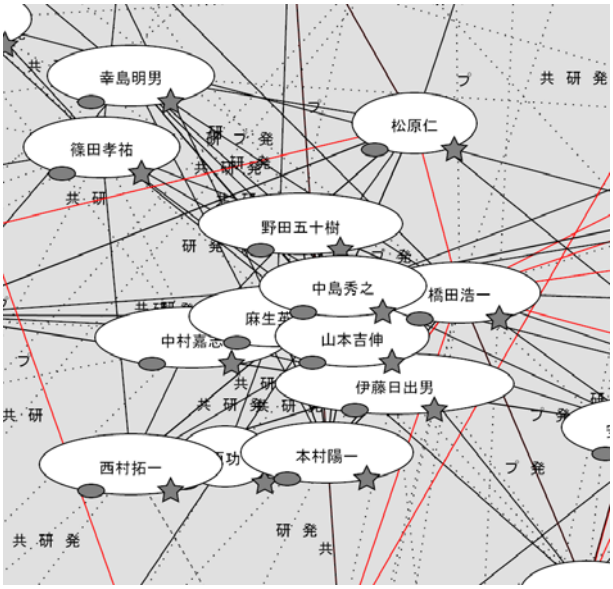

図 1: JSAI で表示した人間関係ネットワーク（拡大図）

表示したネットワークを図 1 に示す．ノード数 266，エッジ数 690[*5] のネットワークである．

ネットワークは，SVG[*6] 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．各ノードには丸印と星印のアイコンがあり，それぞれスケジューリング支援システム，CoBIT による位置情報表示システムと連携している．（各システムの概要は [西村 04] 参照.）エッジは，Simpson 係数 $R(X,Y)$ が閾値（0.7）を越えるノードペア X，Y に対して実線で表示している．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）がそれぞれ 243 本，243 本，92 本，192 本のエッジに付与されており，それらをクリックすると，その判断の根拠となったページへジャンプする．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[*7]．

*5 実線エッジ 284，破線エッジ 262，赤エッジ 144．区別については後述．

*6 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．

*7 Graphviz(`http://www.research.att.com/sw/tools/`

また，$R(X,Y)$ による黒い実線のエッジの他にも，ネットワーク表示によるコミュニケーションの促進となるように，次のような 2 種類のエッジを加えて表示した．

赤エッジ 共起頻度が閾値（100）以上のものについて赤色のエッジを表示する．ヒット件数の多い有名な人のペアが多く含まれ，コミュニティの骨格を表すために有用である．

破線エッジ 各ノードに対してエッジが 3 本以下の場合，閾値をさらに下げて（0.5）破線でエッジを表示する．

JSAI2003 の会場で運用を行うことによって，旧姓の併用の問題，外国人名の問題など，いくつか問題点も明らかになったが，人間関係ネットワークを表示するページへのアクセスも多く，分かりやすく面白いシステムであった，研究者の全体的な関係を理解するのに役立ったなどの声も聞かれた．

3.2 アンケートによる評価

JSAI2003 の後，我々は人間関係ネットワークに関するアンケート調査を行った．調査対象者は，JSAI2003 に参加登録した人の中から選んだ 141 人とした[*8]．CGI によるアンケートシステムを作成し，アンケートの協力をお願いするメールを対象者に送付した．82 名から回答を得，回収率は 58%であった．

アンケートでは，各被験者に対して，スケジューリング支援システムで know リンクを張った／張られた人から 10 人，さらに我々のシステムにおいて共起の閾値つき Simpson 係数 $R(X,Y)$ に応じたルーレット選択で 10 人を抽出し，一人あたり各 20 人の相手との関係を質問した．質問は，一人の相手あたり各 15 問であり，「共著の論文がある（既に公になっているものに限る）」，「同じ研究室や部署など 30 人規模の組織に同時期に所属している，またはしていた」，「同じプロジェクトや委員会に所属している，またはしていた」，「JSAI2003 以外の研究会や国際会議で会ったことがある」などの項目を含む．それぞれ，共著，研究室，プロジェクト，発表の関係の有無を問う意図で設定した質問項目である．

表 5 に，JSAI2003 で表示したネットワークのエッジラベルに対して，アンケートから得た回答を正解とした場合の適合率および再現率を示す．また，表 6 は，抽出した全関係に対する適合率および再現率であり，ネットワーク中にエッジラベルとして表示していないものも含む．

表 4 と比較して，表 5，表 6 は，適合率，再現率ともに低い値になっている．この理由として考えられるのは，

- 回答者が共著やプロジェクトの関係を忘れている，記述もれしているなどの可能性がある．例えば，5 の共著で，システムの出力が誤っていたとされた 10 件 (= 91 件 − 81 件) 中，6 件は実際には共著の関係があった．特に発表関係は，はっきりと覚えていない場合も多いと考えられる．
- プロジェクトの定義としてより広いものを想定しており，再現率が低くなっている．

など，アンケートの回答に関する問題である．しかし，最も大きな原因，特に再現率が低いことに対する原因として考えられるのは，次のような点である．

graphviz/) を使い，ばねモデルによる初期配置を求めている．

*8 JSAI2003 にスケジューリングシステムにメールアドレスを登録した 231 人のうち，スケジューリングシステムにおける know リンクの数と本システムにおけるエッジの数の和が 10 人に達しない 90 人を除いた，141 人全員を対象者とした．アンケート送付は 2003 年 12 月 4 日であり，その後約 2 週間で回答を締め切った．

表 5: JSAI2003 で表示したネットワークにおけるエッジラベルのアンケートによる評価

クラス	適合率	再現率
共著	89.0% (81/91)	32.1% (81/252)
研究室	78.3% (72/92)	18.7% (72/385)
プロジェクト	50.0% (9/18)	3.0% (9/300)
発表	79.5% (35/44)	6.5% (35/538)

表 6: 抽出された全エッジラベルのアンケートによる評価

クラス	適合率	再現率
共著	78.5% (135/172)	53.6% (135/252)
研究室	55.6% (109/198)	28.3% (109/385)
プロジェクト	20.3% (60/296)	20.0% (60/300)
発表	39.9% (222/556)	41.3% (222/538)

- すべての Web ページを網羅的に分析しているわけではない．検索でヒットした上位 5 ページのテキストを分析したものであるため，関係を取り逃す場合もある．
- すべての情報が Web 上にあるわけではない．プロジェクトに関しては，Web 上に情報がないものも多い．また，研究室には過去のメンバーリストを載せておらず，現在のものに書き変わっている場合もある．

そもそも，Web 上にない情報から関係を把握することはできないので，本研究のアプローチが再現率に対して限界があることは明らかである．しかし，関係の強さと併せて用いることで，表 5 で示したように 80%程度の適合率で共著，研究室，発表などの関係を抽出できるということは，学会におけるコミュニケーション支援という目的には有用であろう．

4. まとめ

本稿では，人間関係ネットワークを Web から抽出する方法，および JSAI2003 での図示と評価について述べた．ここで対象としている人間関係ネットワークは，Web 上の情報から抽出した研究者の協働関係のネットワークであるが，例えば，位置情報を用いることで，「出会った」関係を表すネットワーク[*9]，最近流行している blog を用いて誰と誰がよくコミュニケーションしているかというコミュニケーションのネットワークなど，さまざまな形の人間関係ネットワークを取り出すことができる．もちろん，プライバシーの問題には十分気を使う必要があるが，人間関係に基づく情報支援は非常に大きな可能性を秘めていると考えている．

参考文献

[Manning 02] Manning, C. D. and Schütze, H.: *Foundations of statistical natural language processing*, The MIT Press, London (2002)

[Quinlan 93] Quinlan, J. R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann, California (1993)

[西村 04] 西村 拓一, 濱崎 雅弘, 松尾 豊, 大向 一輝, 友部 博教, 武田 英明：2003 年度人工知能学会全国大会支援統合システム, 人工知能学会誌, Vol. 19, No. 1, pp. 43–51 (2004)

*9 我々の解析では，80%程度の適合率と 15%程度の再現率でだれとだれが会ったかを検出できる．