3D4-3

Web 上の情報を用いたアーティスト間のネットワーク抽出

Extracting Artist Network from World Wide Web

金英子[*1] 松尾豊[*2] 石塚満[*1]
YingZi Jin 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
National Institute of Advanced Indutrial Science and Technology

Social network extraction from the Web is receiving much attention recently. This paper proposes a new algorithm to extract social network of artists, which can effectively identify weak social relationships among artists. Four parameters are used in the algorithm. We evaluate our method elaborately, and show the effectiveness of our method. Our system is operated on the Web site for Yokohama triennale 2005.

1. はじめに

日常話題になっているニュースや学会・展示会などのイベント，個人のホームページや Blog など，Web 上には人間の社会的な活動に関する多岐にわたる情報が存在する．Web はある種の人間社会を反映していると言っても過言でない．Web 上の情報は，社会学者の注目も集めており [Wellman04]，Web から社会ネットワークを抽出し分析する研究が盛んに行われている [Mika 05, 原田 03, Kautz 97, Yuta 05]．社会ネットワーク分析とは，行為者をノードとし行為者間の関係をエッジとした関係構造に基づいて物事を分析するすることで，全体や個別の特徴を示す方法論である [安田 97]．松尾らは，特定のコミュニティの人間関係を抽出する手法を提案して，与えられた氏名リストに対して Web から人間関係ネットワークを自動抽出することで，学会などのコミュニティ支援を行っている [松尾 05][Matsuo 06]．本研究では，松尾らの研究の拡張として，現代美術，パフォーマンス，建築などに関するアーティストのネットワークを Web 上から自動的に抽出する．得られたネットワークを用いて，国際展における作品の閲覧を支援することが可能である．

様々な分野のアーティストが参加する国際的な展示会の場合，アーティスト同士の関係性にはばらつきが大きく，関係の強度が大きく異なる特徴がある．このような状況でネットワークを抽出すると，既存の方法では多くの弱い関係性を取り逃して孤立ノードがたくさん出現する．本研究では，客観的にみて弱い関係であってもその人にとって重要な人々を見つけていくことで，ネットワークを構成する手法を提案する．ネットワーク全体に一貫した閾値を用いてエッジを張る従来からの方法に加えて，個々のノードごとにエッジを張る方法をパラメータを用いて調整し，弱い人間関係であっても的確に抽出することができる．なお，本研究は，2005 年 9 月 28 日から 12 月 18 日にわたって横浜市で開催された横浜トリエンナーレ 2005[*1] のアーティストを対象に適用され，得られたネットワークは Web 上で閲覧利用された．

連絡先: 金英子，東京大学大学院 情報理工学系研究科
〒 113-8656 文京区本郷 7-3-1 工学部新 2 号館 111C1 室
TEL: 03-5841-6774, FAX: 03-5841-8570
Email: eiko-kin@mi.ci.i.u-tokyo.ac.jp

*1 www.yokohama2005.jp

2. Web から人間関係ネットワークの抽出

2.1 関係の強さの定義

Web 上から人間関係を同定する基本的な考え方は,「Web における名前の共起の強さは，その 2 人の関係の強さを表している」という仮説に基づいている [松尾 05, Matsuo 06]．ここで，Web における名前の共起とは，同一の Web ページ上に名前が同時に出現することを指す．例えば，研究者の場合では，学会や研究会のプログラム，研究室のメンバーのページ，大学内の教官のメンバーリストなどのページに名前が多く共起するほど，2 人のアクターの間には何らかの社会的関係が強い可能性が高いと推測できる．

Web における人物 x_1 と x_2 の共起の強さを計量するには，上記のような名前の共起頻度を用いる以外に，ダイス係数，相互情報量，コサイン類似度，Jaccard 係数，Simpson 係数などさまざまな尺度がある [Manning 02]．松尾らは，これらを用いて人間関係の共起の強さを計算した場合について評価・考察を行い，Simpson 係数が人の協働関係の強さを表すのに最も適していることを示している [松尾 05]．Simpson 係数は，分母に関して min をとっており，ヒット件数の小さい方から見た距離感を表している．しかし，Simpson 係数は，単独でのヒット件数が非常に少ない人は高い値が出やすいという欠点があるため，次のような閾値 k を設定し，x_1 と x_2 の関係の強さ Rel を求めている．

$$
\begin{aligned}
&Rel(x_1, x_2) \\
&\quad = \quad Simpson(x_1, x_2) \\
&\quad = \quad \begin{cases} \frac{|x_1 \cap x_2|}{min(|x_1|, |x_2|)} & (\text{但し}, |x_1| > k,\ |x_2| > k) \\ 0 & (\text{上記以外}) \end{cases}
\end{aligned}
$$

このように計算した $Simpson(x_1, x_2)$ が閾値（SIM とする）以上であればエッジを張っていくことでネットワークを構成する．つまり，k と SIM という 2 つの閾値をパラメータとして用いてネットワークを構成していることになる．

本論文では，便宜的に名前単独のヒット件数の閾値 k ではなく，共起頻度 $cooc(x_1, x_2) = |x_1 \cap x_2|$ に対して用いられる閾値 CO を用いる．つまり，$cooc(x_1, x_2) \leq CO$ の場合は x_1 と x_2 にエッジを張らない．k を用いる場合には閾値以下のノードは最初から全く考慮しないことに相当するが，CO を用いることでいったんノードとして考慮したあと共起頻度でエッジが

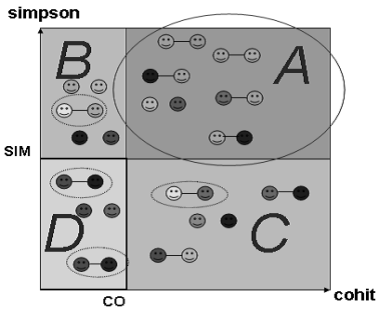


図 1: Simpson 係数と共起頻度による抽出の範囲

調整されると解釈できる．なお，$|x_1 \cap x_2| < \min(|x_1|, |x_2|)$ であるため，k を用いても CO を用いても同様のネットワークを作ることが可能である．

まとめると本論文では，

- SIM: 2つの名前の Simpson 係数の閾値
- CO: 2つの名前の共起頻度の閾値

という2つのパラメータを用いる方法を従来手法として捉える．

2.2 従来手法の問題点

国内の同じ学会（例えば，人工知能学会など）に参加する研究者同士の場合は，同じ分野の研究に従事する研究者がほとんどであるため，Web 上での名前の共起の強さと実際の協働関係の強さには一貫した相関がある．こういった同質なコミュニティ内では，一定の基準を決めることで関係の有無を判断することは可能である．

しかし，国際的に活動するアーティスト同士の場合，関係の強度にばらつきが大きく，一定の基準を設定しづらい．これは国をまたがるアーティスト同士の関係が同じ国のアーティスト同士より弱い関係と捉えられてしまうためである．同様に異なる分野のアーティスト同士の関係は同じ分野のアーティスト同士より弱い関係と計算される傾向になる．また，最近結成されたアーティスト同士は，古くから結成されているアーティスト同士より，Web 上に名前が共起することが少ないので，共起関係が弱くなる．

このように，異質なコミュニティにまたがるネットワークを抽出する際には，アクター間の関係の強さのばらつきが大きい．社会的な活動の結果としての関係は実際に存在するものの，Web 上では共起関係が強く現れないアクター間の関係を，本論文では**弱い社会的関係**と呼ぶ．アーティストの世界では，様々な分野の人々が手を組んで作品を作ったり，作品ごとに協力し合う仲間が異なったりすることがよくあるので，弱い社会的関係であっても，それを適切に捉えることが重要である．

3. 提案手法：弱い社会的関係の抽出

3.1 着想

Simpson 係数と共起頻度という2つの値を2次元で表すと図 1 のようになる．従来手法では，Simpson 係数と共起頻度がそれぞれの閾値 SIM，CO 以上になる A の部分だけを抽出している．B や C の部分は，Simpson 係数と共起頻度のいずれかしか閾値を超えない比較的弱い関係であり，さらに D の部分は，Web 上での共起が多少はあるにしても非常に弱い関係である．国際的な展示会のように，アクター同士の関係の強度に一貫性がない場合には，このような B や C（まれに D）の部分にも社会的弱い関係が存在する．

本節では，社会学におけるネットワーク・クエスチョン[*2] の考えに立ち戻り，それぞれの人にとって重要な人を抽出するネットワーク抽出のアルゴリズムを提案する．

```
Input: all network members name
Output: relational member set for eatch member

X_all = all members name
while X_all ≠ ∅
  x_i ← the ith member in X_all
  X_all ← X_all − {x_i}
  Y_all = all members name; Y_all ← Y_all − {x_i}
  Y = ∅, C1 = ∅, C2 = ∅, C3 = ∅
  while Y_all ≠ ∅
    y_j ← the jth member in Y_all
    Y_all ← Y_all − {y_i}
    if (simpson(x_i, y_j) > SIM ∧ cohit(x_i, y_j) > CO) /*条件 1*/
      then C1 ← {y_i}, ordered by simpson(x, y)
    else if (simpson(x_i, y_j) > SIM ∨ cohit(x_i, y_j) > CO) /*条件 2*/
      then C2 ← {y_i}, ordered by simpson(x, y)
    else if (simpson(x_i, y_j) > 0 ∧ cohit(x_i, y_j) > 0) /*条件 3*/
      then C3 ← {y_i}, ordered by simpson(x, y)
  end
  Y ← C1 /*ルール 1*/
  if (|Y| < M)
    while C2 ≠ ∅, |Y| ≤ M /*ルール 2*/
        c_2k ← the kth strong relation member in C2
        Y ← Y ∪ {c_2k}
    end
    if (|Y| < N)
      while C3 ≠ ∅, |Y| ≤ N /*ルール 3*/
        c_3l ← the lth strong relation member in C3
        Y ← Y ∪ {c_3l}
      end
    end
  end
end
```

図 2: 提案手法のアルゴリズム

3.2 提案手法のアルゴリズム

提案手法の詳細なアルゴリズムを図 2 に示す．その処理は以下の通りである．入力は，ネットワークを構成するアクターの集合 X_{all} で，出力は，アクター間の抽出されたすべての関係である．各アクター $x_i \in X_{all}$ ごとに，関係の強いアクターへのエッジを構成する．自分自身以外のすべてのアクター（Y_{all} で表される）との関係を調べて，次の各条件にしたがって自分と関係の強さによって3つのクラス（$C1(x_i)$ から $C3(x_i)$ まで）に分ける．このクラスは，後でルールを用いてエッジを張る際に用いる．

C1($\mathbf{x_i}$) x_i との Simpson 係数および共起頻度が閾値以上であるアクターのクラス

C2($\mathbf{x_i}$) x_i との Simpson 係数，共起頻度のいずれかが閾値以上であるアクターのクラス

C3($\mathbf{x_i}$) Simpson 係数と共起頻度のいずれも閾値未満（値は 0 より大）であるアクターのクラス

なお，共起頻度が 0 のものはいずれのクラスにも属さない．

*2 従来社会学で人のネットワークを抽出するために用いられる代表的な方法で，例えば，「あなたが過去半年のあいだに，あなたにとって重要なことを話しあった人々は誰でしたか？」というネットワーク質問をそれぞれの人の対して行う．このような質問により，個人のもつネットワークや関係性の分析を行うことができる [安田 97]. しかし，多くの人に対して定期的にこのようなアンケートを行うことは難しい．

ここで，各クラスを図 1 に対応させて考えると，クラス $C1(x_i)$ は A 部分に，$C2(x_i)$ は B ないしは C 部分に相当する．$C3(x_i)$ は D 部分であり，最も優先順位の低いクラスである．

次に，以下のルールにしたがって x_i からのエッジを生成する．

ルール 1 x_i と $C1(x_i)$ に属するアクターをエッジでつなぐ．

ルール 2 ルール 1 の結果，x_i のエッジの数が M より少ない場合，$C2(x_i)$ に属するアクターをエッジ数が最大で M に達するまでエッジでつなぐ．

ルール 3 ルール 2 の結果，x_i のエッジの数が N より少ない場合，$C3(x_i)$ に属するアクターをエッジ数が最大で N に達するまでエッジでつなぐ．

つまり，提案手法では CO，SIM，M，N という 4 つのパラメータを考慮に入れる．なお，$M = 0$，$N = 0$ とした場合には，従来手法と同じネットワークが得られる．

4. 提案手法の評価

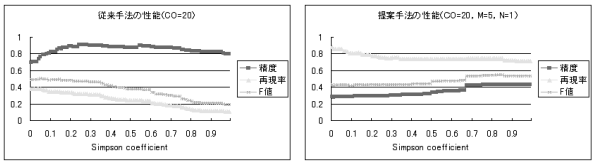
提案手法の有効性を示すために，テストデータを作り，従来手法と提案手法による性能を比較した．テストデータは，「同グループ」関係のアーティストのペアを 146 個と関係がないペア 854 個，合計 1000 個のペアから構成した．ここでは，適合率，再現率および F 値による評価を行う．F 値は次の式で定義される．

$$F\text{ 値} = \frac{2 \times (\text{適合率}) \times (\text{再現率})}{(\text{適合率}) + (\text{再現率})}$$

表 1 は，従来手法を用いて，テストデータに対して最大の適合率，再現率，および F 値が得られるときの各値とパラメータ SIM および CO の値を示している[*3]．最大の再現率を得るには SIM や CO を下限に設定すればよいので，$SIM = 0$，$CO = 0$ のとき再現率は 100%である．しかし，適合率は 14.6%と非常に低い．逆に適合率を最大にするには，共起の高いペアだけを選べばよいので，$SIM = 0.24$，$CO = 30$ と高く設定したときに，適合率は最大の 92.9%，再現率は 26.7%となる．両方のバランスを取る最も良いパラメータは，$SIM = 0.05$，$CO = 20$ であり，このとき F 値が最大の 0.50 となる．

一方，提案手法における結果を表 2 に示す．提案手法では 4 つのパラメータがあり，SIM と CO が表 1 と同じ値であるとした場合でも，M と N を適切に調整することで F 値が上がることを示している．全ての値を適切に調整すると，F 値は最大で 0.55 になる．

次に，この結果がどの程度ロバストかを，パラメータを変化させながら示す．図 4(a) と図 4(b) は，従来手法と提案手法に対し，SIM を変化させたときに適合率，再現率，F 値がどのように変化するかを示している．(a) からは，2 節で従来手法の問題点として述べたように，SIM を高く設定すると実際に存在する弱い社会的関係を取り逃がし再現率が低くなり，逆に閾値を低く設定すると関係のないエッジが多く生成され適合率が下がる．一方の (b) では，$M = 5$，$N = 1$ のパラメータを加えることで，再現率が従来手法より高い値を保ち，SIM によらず F 値が比較的安定していることが分かる．



(a) 従来手法　　(b) 提案手法

図 4: SIM を変化させたときの適合率，再現率，F値の変化

これらのパラメータを用いて構築されたネットワークを比較してみると，従来手法では図 3(a) に示すように多くの孤立ノードが生成されるのに対し，(b) の提案手法ではノードが孤立していない．

ここで示した結果は，提案手法が従来手法より有効であることを示している．しかし，提案手法では従来手法よりアルゴリズムの自由なパラメータが増えているので，この結果はある程度当然の結果であるといえるかもしれない．しかし，ネットワーク全体に一貫した閾値（SIM，CO）を設定することと，各ノードから見た閾値（M，N）を設定することは，これまでにもネットワークを抽出する多くの研究で，無意識に混在されて用いられていた．本論文でこれをパラメータとして捉え，その性能評価となるデータを示せたことは，少なくとも Web や電子メール等から社会ネットワークを抽出するさまざまな研究に，重要な知見と示唆を与えるものである．

5. 関連研究との位置づけ

Web を情報源として，人間関係のネットワークを抽出する研究には，本論文で説明した松尾らの研究の他にも，Mika の研究 [Mika 05] や原田らの研究 [原田 03]，Referral Web[Kautz 97] などがある．Mika は，FOAF データを抽出する手法として，松尾らの手法とほぼ同様に Web 上の共起ヒット件数を用いて，共起関係の強さを判断しているが，Jaccard 係数を用いている点と関係にラベルをつけていない点が異なる．原田らは，ユーザが入力するトピックに関して，Web 上から上位にヒットする Web ページを取得して人名リストを抽出し，人名が共起する距離に基づいて関係の強さを計算し，トピックに関連する人間の関係を抽出している．Referral Web[Kautz 97] では，個人のもつネットワークを抽出するために，その人の氏名で検索した上位のページに含まれる氏名リストと Jaccard 係数を用いて関係性を調べ，さらに氏名リストをそれぞれクエリとして検索することを繰り返すことで，その人を中心とするネットワークを抽出している．

これらの手法は，Web 上の社会的な関係の強度をどのように計量化するかという点で異なる手法であるが，いずれも同質なコミュニティ内で一貫的に決めた関係の基準によって関係のあるなしを判断している．そのため，ばらつきが大きいアクター間の関係性を調べる時に，弱い社会的関係を抽出できない問題点がある．本研究で提案しているネットワーク抽出手法は，こういった方法と社会学におけるネットワーク・クエスチョンという典型的な方法の両方を統合するものであり，Web からの人間関係ネットワーク抽出のより汎用な方法である．

6. まとめ

本研究では，アーティストのような弱い社会的関係であっても，Web 上から適切にネットワークを抽出する手法について提案し，評価実験により提案手法の有効性を示した．本手法で得られたアーティスト間のネットワークは，横浜トリエンナーレの開催期間中に Web サイト上で運用された．

*3 各パラメータは，SIM は 0 から 1 まで 0.01 ずつ，CO は 0 から 60 まで 5 ずつ，M は 0 から 5 まで 1 ずつ，N は 0 から 4 まで 1 ずつ変化させて検証した．以下でも同様である．

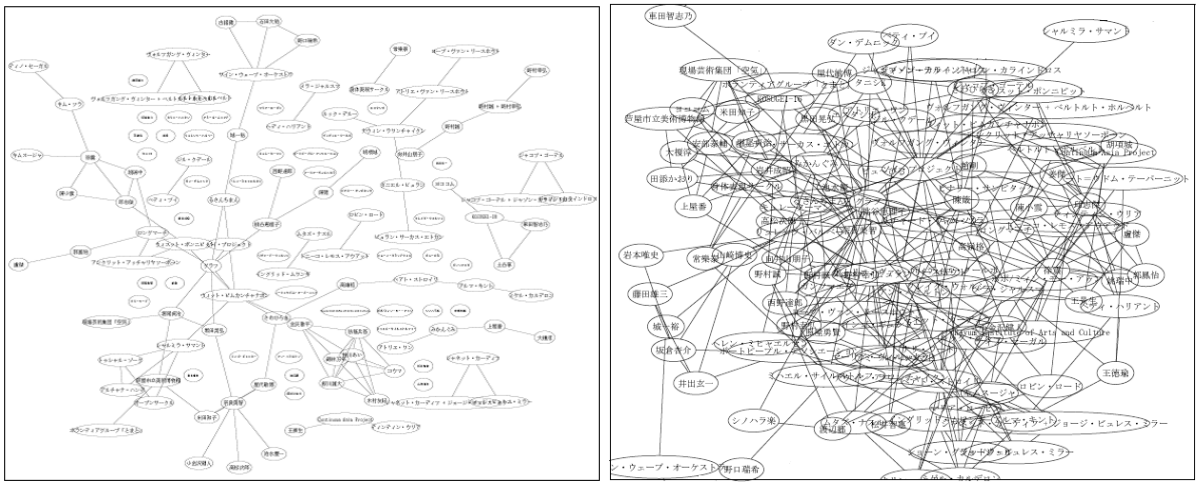
表 1: 従来手法の適合率，再現率，F 値と各パラメータの値

性能	SIM	CO	適合率	再現率	F 値	抽出されたエッジ数*	正解エッジ数*
(a) 適合率が最大の場合	0.24	30	92.9%	26.7%	0.41	42 (42,0,0)	39 (39,0,0)
(b) 再現率が最大の場合	0	0	14.6%	100%	0.25	1000 (1000,0,0)	146 (146,0,0)
(c)F 値が最大の場合	0.05	20	76.4%	37.7%	0.50	72 (72,0,0)	55 (55,0,0)

∗: 括弧内は，それぞれ $C1$, $C2$, $C2$ の各クラスからのエッジ数を示している．

表 2: 提案手法の適合率，再現率，F 値と各パラメータの値

	SIM	CO	M	N	適合率	再現率	F 値	抽出されたエッジ数	正解エッジ数
(a) の場合	0.24	30	3	2	34.4%	65.1%	0.45	277 (42,227,8)	95 (39,54,2)
(b) の場合	0	0	0	0	14.6%	100%	0.25	1000 (1000,0,0)	146 (146,0,0)
(c) の場合	0.05	20	1	0	55.4%	49.3%	0.52	130 (72,58,0)	72 (55,17,0)
F 値が最大の場合	0.82	20	5	1	43.4%	74.0%	0.55	249 (23,212,14)	108 (19,84,5)



(a) 従来手法
（$SIM = 0.24$，$CO = 30$）

(b) 提案手法
（$SIM = 0.24, CO = 30, M = 3, N = 2$）

図 3: 抽出されるネットワークの違い

今後は，さまざま目的のネットワークにおいて，各パラメータはどのようなものが適切であるのか，またエッジのラベルを判別するモジュールの性質がどのようにネットワーク抽出の精度の向上につながるかといった研究をさらに進めていきたいと考えている．

参考文献

[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, 情報処理学会研究会報告，DBS-130/FI-71(2003)

[Kautz 97] H. Kautz, B. Selman and M. Shah: The Hidden Web, AI magazine, Vol.18, No.2, pp.27-35(1997)

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

[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)

[Matsuo 06] Y. Matsuo, J. Mori, M. Hamasaki, Keisuke Ishida, Takuichi Nishimura, Hideaki Takeda, Koiti Hasida, and Mitsuru Ishizuka: POLYHPONET: An Advanced Social Network Extraction System from the Web, In Proc. WWW 2006(2006)

[Mika 05] P. Mika, Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks, Journal of Web Semantics, 3:2(2005)

[Miki 05] Takeru Miki, Saeko Nomura and Toru Ishida, Semantic Web Link Analysis to Discover social Relationship in academic communities, in Proc. SAINT 2005(2005)

[安田 97] 安田雪: 社会ネットワーク分析 －何が行為を決定するか－, 新曜社 (1997)

[Yuta 05] 湯田 聴夫, 藤原 義久 SNS における人のネットワーク構造 〜その地平線の超え方〜, Web が生み出す関係構造と社会ネットワーク分析ワークショップ, 社会情報学フェア (2005)

[Wellman04] Barry Wellman: The Global Village: Internet and Community, Idea&s - The Arts & Science Review, University of Toronto, 1(1): 26-30(2004)