3F1-02

Web 上の情報からの大規模人間関係ネットワークの構築と分析

Large-scale Social Network Extraction from the Web

浅田 洋平[*1] 友部 博教[*2] 松尾 豊[*3] 石塚 満[*1]
Yohei Asada Hironori Tomobe 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, The University of Tokyo

[*2]名古屋大学大学院 情報科学研究科
Graduate School of Information Science, Nagoya University

[*3]産業技術総合研究所 サイバーアシスト研究センター
Cyber Assist Research Center, National Institute of Advanced Industrial Science and Technology

A social network is useful to show an overview of human relations in a community. We can find a person, with the help of a social network, who is familiar with the person we want to get acquainted with. In our previous work, we proposed a method to measure relation strength between two individuals based on the number of retrieved Web pages that contain both of their names. However, this method overloads a search engine because it issues a lot of unneccesary queries where the search engine reports no Web pages. This nature of the method leads to difficulty when we extract a large-scale social network. We improve the previous method to reduce the unnecessary queries and facilitate the extraction of a large-scale social network. Our experiment shows that the proposed method reduces 85% of the queries preserving the quality of the network compared with the former method.

1. はじめに

近年，研究者が自分の Web ページを持ち，各自の研究成果・研究実績を Web 上で公開する，ということは当たり前のように行われている．また，全国大会や研究会のプログラムや，論文誌に掲載された論文の題目なども，インターネット上に掲載されることが多くなってきた．こうしたことを背景に，我々は，Web 上の情報から人間関係ネットワークを抽出する手法を研究してきた．

人間関係ネットワークにより，自分が他の研究者とどのような関係にあるのかを知ることは，研究者にとって専門分野における自分の位置を把握するだけでなく，共同研究の相手を探したり，新しい研究の種を探したりする際にも役に立つと考えられる．

しかし，周囲の人とだけ付き合っていたのでは不十分である．一見自分の研究とは関係のなさそうな分野にこそ，新しい研究の種は転がっているかもしれないからである．したがって，大規模な人間関係ネットワークを抽出し，誰がどのような研究を行っているのか，ということに関する概観を得ることは，研究者にとって非常に有益であると考えられる．

本稿では，Web 上の情報を用いて大規模人間関係ネットワークを抽出する手法を提案し，それによって日本の情報系研究者のネットワークを構築する．

2. 人間関係ネットワークの抽出

人間関係を抽出する際の情報源としては，

1. E-mail のログ
2. Web ページのハイパーリンク
3. Web における名前の共起

などがある．

E-mail のログを用いれば，個人的な人間関係抽出が可能だが，プライバシーの問題があるため，小さなコミュニティでの適用にとどまらざるを得ない．

ユーザの Web ページのハイパーリンクを用いて人間関係を抽出するものとして，SocialPathFinder がある [4]．ユーザ個人の Web ページにおけるハイパーリンクを用いることで，個人的な人間関係を抽出することができるが，Web ページを持っていないユーザとの関係は抽出できないという問題点もある．

Web における名前の共起を用いて人間関係を抽出するものとして，Referral Web がある [2, 3]．システムは，ユーザから名前を受け取り，Web 上での名前の共起関係にもとづいて，ユーザを中心とする social network を表示する．我々の手法と近いが，前もって名前のリストを用意しておかない点が異なる．

また，ユーザが自分で知り合いを明示的に記述することによってネットワークを構築する枠組みとして，FOAF がある [1]．しかし，ユーザにとって全ての知り合いを記述するのは困難であるという問題点があるし，知り合いが知り合いを記述してくれなければネットワークは広がらない．

これらの研究では，自分の周辺のネットワークの表示に重点がおかれているようであるが，我々の手法は，前もって名前のリストを用意しておくことで，Web 上の情報のみを用いて，自分の周辺のネットワークのみならず，そのコミュニティの人間関係の概観を表示できる，という点が特徴である．

3. 従来手法

我々は，Web 上の情報における共起の度合いに着目して人間関係を抽出する手法を研究してきた [6]．本節では Web からの人間関係ネットワーク抽出の従来手法を紹介し，その問題点について述べる．

連絡先: 浅田 洋平，東京大学大学院 情報理工学系研究科 石塚研究室，〒 113-8656 東京都文京区本郷 7-3-1，Tel. 03-5841-6774，E-mail: asadayo@miv.t.u-tokyo.ac.jp

3.1 基本的な考え方

我々の手法の最も基本的な考え方は，

「Web ページにおいてよく名前が共起している二人の間には，何らかの関係がある」

ということである．

例えば，「石塚満」「浅田洋平」という二人の間に関係があるかどうかを調べることを考える．「石塚満」「浅田洋平」という二人の間に関係があるかどうかを調べるには，「石塚満」「浅田洋平」の二人の名前が Web 上でどの程度共起 [*1] しているかを調べればよい．そのためには，「石塚満」「浅田洋平」が共起している Web ページが何ページあるのかを知ることが必要になる．

すべての Web ページの中から「石塚満」「浅田洋平」が共起しているページが何ページあるかを調べることは非常に困難であるので，「石塚満 and 浅田洋平」をクエリとして検索エンジンに与え，そのヒット件数を，「石塚満」「浅田洋平」が共起している Web ページ数とみなす．

3.2 関係の強さの定義

共起ページ数を用いて，X, Y という二人の関係 $R(X, Y)$ を式 (1) により定義する．ただし，名前「X」「Y」の単独ヒット件数を「$|X|$」「$|Y|$」，「X and Y」のヒット件数を「$|X \cap Y|$」と書くことにする．

$$R(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)} \quad (1)$$

これは，overlap 係数と呼ばれるものである．式 (1) について定性的な考察を加えると，$|X|, |Y|$ の小さいほうを分母としているため，単独ヒット件数の少ない人との関係が強くなりやすいという傾向が分かる．極端な例では $|X| = 1, |Y| = 100, |X \cap Y| = 1$ の場合を考えると，1 ページしか共起していないにもかかわらず，$R(X, Y)$ は最大の 1.0 となってしまう．

このため，実際には，$|X|, |Y|$ のどちらかが閾値以下の場合には，その二人の関係は誤差のうちとして，ないものとみなしている [6]．

3.3 従来手法の処理の流れ

従来手法の処理の流れは，次のようなものである．

Step1 名前のリストを用意する

Step2 全ての名前について，その名前を含む Web ページ数を調べる

Step3 リスト中の全ての名前の組み合わせについて，それらが共起する Web ページ数を調べる

Step4 リスト中の全ての名前の組み合わせについて，それらの関係の強さを示す overlap 係数を計算する

3.4 従来手法の問題点

従来手法には，次のような二つの問題点がある．

検索エンジンに対する負荷の問題 従来手法では，作成したいネットワークを構成するメンバーの数を n とすると，全てのメンバーのペアの組み合わせについてその二人が共起する Web ページ数を調べるので，${}_nC_2$ 回の検索を行うことになる．これは，n が大きくなるにつれて，ほぼ n^2 のオーダで大きくなるため，大規模なネットワークを作成する際には，検索エンジンに対する負荷が非常に大きくなり，事実上，ネットワークの作成が不可能になる．

共起行列のスパースネスの問題 従来手法では，上記のように全てのペアに対して検索をし，共起頻度を調べているが，一人の人が必ずしもそのほかの全ての人と関係があるわけではなく，また，overlap 係数が閾値以下の弱い関係は最終的にネットワークには表示されないため，共起行列は非常にスパースなものとなる．したがって，全てのペアについて共起件数を調べることは，非常に効率の悪いアルゴリズムであると言える．

*1 ただし，ここでは，同一の Web ページにおいて二人の名前が出現していることを「共起」と呼ぶことにする．

4. 提案手法

本節では，従来手法の問題点を解決する手法を提案する．

4.1 提案手法の考え方

従来手法の問題点は，リスト中の全ての名前の組み合わせに対して，それらが共起している Web ページ数を調べていたことであった．

例えば，ある名前をクエリとして検索エンジンで検索した結果の上位のページを良く見てみると，そこにはその人と関係の強い人の名前が出てくる．このことを利用すれば，強い関係のありそうな人と共起関係だけを調べることができると考えられる．

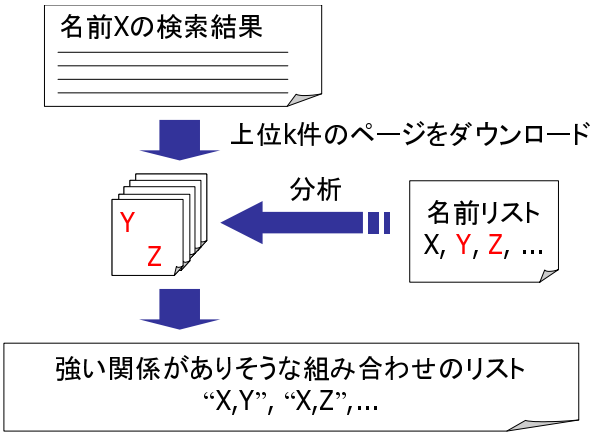

図 1: 提案手法の概要

そこで，図 1 のように，ある名前の検索結果のページから，上位 k 件のページについて，その内容を分析し，リスト中の名前が含まれていないかを調べる．そして，この上位 k 件のページに含まれる名前との間には強い関係がありそうだと推測し，以後，これらの名前との共起のみを調べることにする．これにより，検索エンジンに対する負荷を減らすことができると考えられる．

4.2 提案手法の処理の流れ

提案手法の処理の流れは，次のようになる．

Step1 名前のリストを用意する

Step2 全ての名前について，その名前を Web ページ数を調べる

Step2' 単独の名前の検索結果における上位 k ページの内容を分析し，リスト中に含まれている名前があれば，その組み合わせは関係が強そうだと推測し，リストアップする

Step3 リストアップした組み合わせに対してのみ，それらが共起する Web ページ数を調べる

Step4 overlap 係数を計算する

5. 評価実験

前節で提案した手法の有効性を検証するために，同じネットワークを従来手法，提案手法の二つの方法で作成し，その結果を比較する実験を行った．本節では，実験とその結果に対する評価を行う．実験に用いたネットワークのデータを表 1 に示す．

表 1: 実験に用いたネットワークのデータ

ノード	JSAI2003 の参加者
ノード数	503

提案手法では，単独の名前の検索結果の上位 20 件のページを分析した ($k = 20$).

5.1 評価方法

従来手法では，名前リストから考えられる全ての名前のペアについて，その共起関係を調べていた．

これに対して，提案手法は，単独の名前を含む Web ページの上位 20 件のページを分析し，それらの中で共起している名前についてのみ，強い関係がある可能性のあるペアであるとみなして実際に共起関係を調べることにより，検索エンジンに対する負荷を少なくするというものである．

すなわち，従来手法では全ての関係を抽出できるが，提案手法では関係を見落としてしまう可能性がある．

したがって，評価は，提案手法を用いることで，従来手法に対して，

1. どれだけ検索エンジンに対する負荷が減ったか
2. どの程度の関係を抽出できているのか

の 2 点に対して行った．

後者の点については，提案手法は，上位 20 件のページに出現する名前との共起しか調べていないので，弱い関係を抽出することは難しくなる．そこで，overlap 係数に閾値 $threshold$ をもうけ，その閾値以下の関係をないものとみなした場合に，関係を何%抽出できたかを調べる目的で，式 (2) のように $Coverage$ を定義した．

$$Coverage = \frac{\text{提案手法による } threshold \text{ 以上の関係の数}}{\text{従来手法による } threshold \text{ 以上の関係の数}} \quad (2)$$

この $Coverage$ を調べることで，提案手法によりどの程度の関係を何%抽出できたのかということを知ることができる．

5.2 実験結果

まず，従来手法と提案手法のそれぞれの手法を用いた場合において，共起頻度を調べるために検索エンジンに与えるクエリの数を表 2 に示す．

表 2: クエリ数の比較

従来手法によるクエリ数	126253
提案手法によるクエリ数	19182

次に，提案手法によってどの程度の強さの関係を抽出できたのかを調べる目的で，すべての関係について提案手法と従来手法のそれぞれによって overlap 係数を求め，その相関関係を調べた．その結果を図 2 に示す．

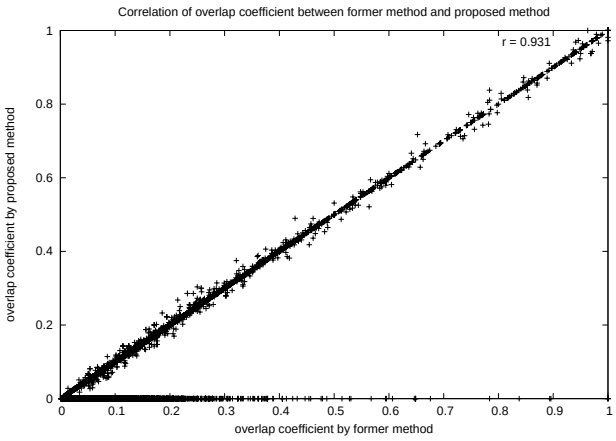

図 2: 従来手法と提案手法による overlap 係数の相関

また，提案手法によってどの程度の関係を何%抽出できたかを調べるため，overlap 係数の閾値を変えた場合の Coverage の値を式 (2) によって求めた．その結果を図 3 に示す．

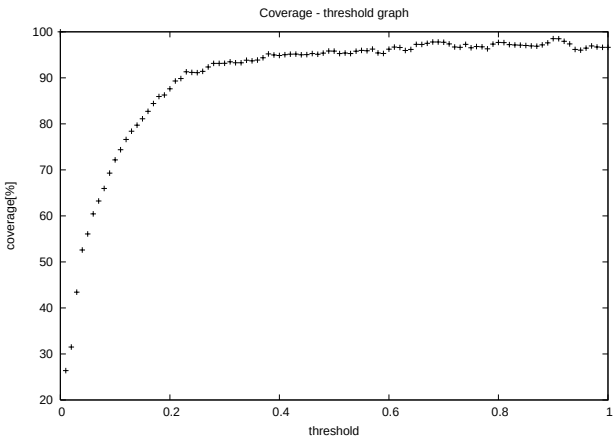

図 3: $Coverage$ と overlap 係数の閾値 ($threshold$) の関係

5.3 考察

表 2 から，提案手法を用いることで，検索エンジンに与えるクエリの数が約 85%減っているということが分かる．このことから，提案手法を用いることで検索エンジンに対する負荷が大幅に減っていることが分かる．

次に，図 2 を見ていただきたい．図 2 における各ドットは，関係を示している．したがって，理想的には全ての関係は

$(0,0),(1,1)$ を結ぶ直線上に並ぶはずであり，x 軸に張り付いているドットが，提案手法によって抽出できなかった関係を示していることになる（ちなみに，従来手法，提案手法ともに抽出された関係であっても正確には $(0,0),(1,1)$ を結ぶ直線上にないものがあるのは，検索エンジンのデータベースが日々更新されていることによるものである）．図 2 から，傾向として，overlap 係数が大きくなるにしたがって，すなわち強い関係ほど，提案手法でもれなく抽出できていることが分かる．ピアソンの積率相関係数を求めると，$r = 0.931$ であった．

また，図 3 から，overlap 係数が 0.2 以上の関係については約 88%，0.3 以上の関係については，約 93%の精度で抽出できていることが分かる．現在，人間関係ネットワークを表示する際には，overlap 係数が 0.5 以上の関係を中心に，エッジの数が少ない場合に 0.2 以上の関係も表示するようにしている．すなわち，overlap 係数 0.2 以上の関係を抽出できていれば，ネットワークの表示には十分である．

以上から分かったことをまとめると，提案手法を用いると，従来手法に比べて

1. 検索エンジンに対する負荷が 85%減少し，
2. overlap 係数 0.2 以上の関係については約 88%抽出できる

となる．

これらのことから，提案手法の有効性を示すことができた．

5.4 大規模人間関係ネットワークの抽出

提案手法を用いて，大規模人間関係ネットワークを抽出した．名前リストは，Web 上の研究者データベース ReaD [*2]から，情報系研究者をリストアップすることで作成した．ネットワークのデータを表 3 に示す．

表 3: 情報系研究者ネットワークのデータ

ノード	ReaD 研究開発支援ディレクトリの「情報工学」「複合領域 - 情報科学」の研究者
ノード数	2879

このネットワークの一部を図示したものが図 4 である．

ちなみに，このネットワークを作成する際に，共起頻度を調べるために検索エンジンに与えたクエリは 137967 個であった．従来手法を用いた場合には 4142881 個のクエリを与えなければならなかったはずであり，ネットワークが大規模になるほど提案手法が有効になることが分かる．

6. おわりに

本稿の提案手法を用いることで，大規模な人間関係ネットワークを作成，表示することが可能になった．しかし，図 4 から分かるように，大規模ネットワークでは，各人が複雑に関係しあっており，全体像を見ただけでは誰が誰と関係があるのかひと目で分からないという新たな問題点が分かってきた．人間関係ネットワークを効率よく利用できるように様々な工夫が必要である．

今後次のように研究を進め，ネットワークから自分と興味の似ている人や，新しい研究の種を見つけ出せるようなアプリケーションを開発したいと考えている．

*2 ReaD 研究開発支援ディレクトリ: http://read.jst.go.jp/

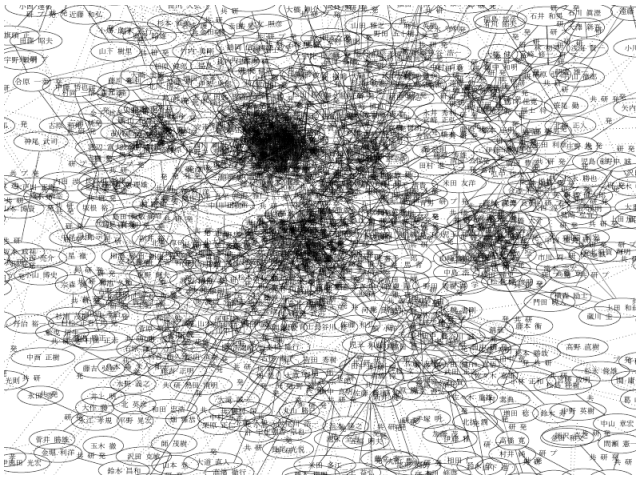

図 4: 情報系研究者ネットワークの一部

ノードに対するラベルの検討 現在，各ノードの持つ情報は名前のみである．そこで，ノードに研究分野などの情報を持たせることで，幅広い研究をしている人や，ある専門分野に関するエキスパートなど，様々な人探しができるようになると考えている．

ネットワークの分析 Social Network に関する研究では，ネットワークの分析が盛んに行われている．例えば，学生の友達関係のネットワークを分析することで，同じ興味を持つものが友達関係になりやすい，という傾向が分かったという報告がなされている [5]．研究者のネットワークについても，様々な分析を行ってみたいと考えている．例えば，クラスタ分析により，研究者のコミュニティを抽出できるのではないかと考えている．

参考文献

[1] Foaf: http://xmlns.com/foaf/0.1/

[2] Henry Kautz, Bart Selman, and Mehul Shah. The Hidden Web. AI Magazine, 18(2), pp.27-36, 1997.

[3] Henry Kautz, Bart Selman, and Mehul Shah. ReferralWeb: Combining social networks and collaborative filtering. Communications of the ACM, vol.40, no.3, 1997.

[4] Hiroaki Ogata, Takayuki Fukui, Yoneo Yano. SocialPathFinder: Computer Supported Exploration of Social Networks on WWW. ICCE99, vol.2, pp.768-771, 1999.

[5] Lada A. Adamic, Orkut Buyukkokten, and Eytan Adar. A social network caught in the Web. “First Monday” http://www.firstmonday.dk/, vol.8, no.6, 2003.

[6] 松尾豊，友部博教，橋田浩一，石塚満．Web からの人間関係ネットワークの抽出と情報支援．人工知能学会全国大会，1F1-02, 2003.