

Webからの研究者ネットワーク抽出の大規模化

Increasing Scalability of Researcher Network Extraction from the Web

浅田 洋平
Yohei Asada

東京大学大学院 情報理工学系研究科
School of Information Science and Technology, University of Tokyo
asadayo@miv.t.u-tokyo.ac.jp

松尾 豊
Yutaka Matsuo

独立行政法人 産業技術総合研究所
National Institute of Advanced Science and Technology
y.matsuo@carc.aist.go.jp, <http://www.carc.aist.go.jp/~y.matsuo/homepage/top.htm>

石塚 満
Mitsuru Ishizuka

東京大学大学院 情報理工学系研究科
School of Information Science and Technology, University of Tokyo
ishizuka@miv.t.u-tokyo.ac.jp, <http://www.miv.t.u-tokyo.ac.jp/~ishizuka/>

keywords: Web mining, search engine, cooccurrence, social network, scalability

Summary

Social networks, which describe relations among people or organizations as a network, have recently attracted attention. With the help of a social network, we can analyze the structure of a community and thereby promote efficient communications within it. We investigate the problem of extracting a network of researchers from the Web, to assist efficient cooperation among researchers. Our method uses a search engine to get the cooccurrences of names of two researchers and calculates the strength of the relation between them. Then we label the relation by analyzing the Web pages in which these two names cooccur. Research on social network extraction using search engines as ours, is attracting attention in Japan as well as abroad. However, the former approaches issue too many queries to search engines to extract a large-scale network. In this paper, we propose a method to filter superfluous queries and facilitates the extraction of large-scale networks. By this method we are able to extract a network of around 3000-nodes. Our experimental results show that the proposed method reduces the number of queries significantly while preserving the quality of the network as compared to former methods.

1. はじめに

近年、研究に関するさまざまな情報が Web 上に公開されるようになった。大学や研究所をはじめとする各研究機関では、研究に関する情報を整理し分かりやすく紹介することに努めており、各種の論文や文献のオンライン化、データベース化も進んでいる。さらに、競争的研究資金に関する採択課題や現在進行中のプロジェクトに関する参加者や発表文献の情報なども公開されることが多くなってきた。

一方で、研究成果を社会に還元するために、産学官の連携が重要視されており、研究者がいかに地域や企業と協力していく体制を作るかは大きな課題である [吉川 02]。また、ロボットやバイオインフォマティクス、ナノテクノロジーなど、複数の関連分野が協力して研究を進める必要のある融合領域、分野横断研究の必要性も高まっている。例えば、昨今では、中越地震やインド洋の津波が大きな被害をもたらしたが、災害救助に関する研究を実

際に社会に役立てるには、地域住民や行政、企業と研究者が密接に連携する必要がある。

このような背景から、著者らは、研究者の効果的な協働関係の構築のために、研究に関する情報を Web から抽出し、研究者ネットワークおよび研究者に関連する情報を抽出する研究を行ってきた [松尾 05, 友部 04, 安田 04, Mori 04]。基本的には、Web 上のページにおける名前の共起関係を検索エンジンを用いて調べ、テキスト処理を組み合わせることで、研究者の関係の強さおよびその種類を特定する。検索エンジンを効果的に使って社会ネットワークを得ようとする研究は、国内外でも注目されつつあり [Mika 04, 原田 03]、今後、Web にある大量の情報を統合することで高次の情報を抽出し、ユーザの文脈に応じた情報支援を行う方向性は大きな潜在的な可能性を秘めていると考えられる [藤井 04]。

Web 全体を大きなデータベースとみなし情報統合を行うアプローチは、さまざまな情報が電子化され Web からアクセスできるようになったこと、さらに検索エンジン

が広く利用可能になったことにより有効性を増している。しかし、一般的なコーパスを用いたテキスト処理と最も大きく異なる点は、検索エンジンをいかに効率的に使うかにある。検索エンジンが広く利用可能になったと言っても、イントラの内部に持つ DB のように自由に SQL 文を書けるわけではない。例えば、Google API では、1 アカウントあたり 1 日 1000 件の限定のもとに Google に検索クエリーを投げることができる*1。検索エンジンに対する負荷を下げることは、今後、検索エンジンを利用した情報統合を行う上で避けては通れない重要な課題である。通信ネットワークのレイヤーでは、帯域を有効に使うためのパケットの通信方式に関する多くの研究があるが、より高次の意味処理に近いレイヤーでも、Web 上のリソースを多くの人々が効率的に使えるようにアルゴリズムを工夫するという視点が、今後、必要性を増してくるであろう。

本論文では、検索エンジンを用いて、研究者の協働関係ネットワークの抽出を行う処理に対して、大規模化を可能にするアルゴリズムを提案する。従来、Web から社会ネットワークを抽出する手法は、大きく 2 つのアプローチがあったが、それを融合するひとつの効率的なアルゴリズムを提案している。それにより、従来手法で 500 人程度のネットワークを抽出するのと同程度の検索負荷で、提案手法では 3000 人程度のネットワークを抽出することを可能にしている。

本論文は次のように構成される。2 節では、本研究のもとになった研究者ネットワークの抽出アルゴリズムと関連研究について述べる。3 節では、アルゴリズムを提案し、4 節で提案手法を用いて抽出した大規模なネットワークの具体例を示し、5 節で提案手法の評価を行う。6 節で他の研究との関係と本研究の位置づけなどについて議論し、7 節でまとめを述べる。

2. 関連研究とその問題点

著者らは、Web における氏名の共起に基づいて研究者ネットワークを抽出する手法を提案している [松尾 05]。その概要は次の通りである。

- (1) 研究者の氏名と所属のリストを与える。このリストは、研究者ネットワークを抽出したいコミュニティのメンバーが入ったものである。例えば、人工知能の研究者ネットワークを得たいのであれば、人工知能学会の全国大会の著者、共著者リストなどを用いる。
- (2) 各研究者に対して、氏名を検索エンジンのクエリーとして検索ヒット数を求める。氏名を X とすると、 X を含む Web ページの集合を D_X 、ヒット件数を $|D_X|$ と表すことにする。実際には「氏名 AND 所属」をクエリーとする検索を行うことで、同姓同名の別人に関する情報を除去している。

- (3) 研究者 2 人のペアに対して、2 つの氏名の Web 上での共起回数を求める。2 つの名前を並べて検索エンジンのクエリーとし（例えば「浅田洋平 松尾豊」）、その検索ヒット数を得る。氏名を X, Y とすると、 X, Y を含む Web ページの集合を $D_{X \cap Y}$ 、ヒット件数を $|D_{X \cap Y}|$ とする。
- (4) $|D_X|, |D_Y|, |D_{X \cap Y}|$ から、2 人の氏名の共起の強さ $R(X, Y)$ を求める。例えば、overlap 係数 $R(X, Y) = |D_{X \cap Y}| / \min(|D_X|, |D_Y|)$ を用いる。ただし、overlap 係数を用いた場合、単独ヒット件数が極端に少ない人との関係の強さを正確に測れないという問題点がある。極端な例として、 $|D_X| = 1, |D_Y| = 100, |D_{X \cap Y}| = 1$ 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、overlap 係数は最大の 1 となる。このような現象が問題となるのは、Web 上に情報がほとんどない学生などの研究者 (X) が、全国大会の参加者リストなどのページ上で多数の研究者 (Y) と共起し、それら全ての研究者との overlap 係数が 1 となってしまう場合などである。この問題を解決するため、[松尾 05] では、式 (1) のような閾値付き overlap 係数を用いている。

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

閾値付き overlap 係数では、 $|D_X|, |D_Y|$ のどちらかが閾値 k ([松尾 05] では 30) 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。[松尾 05] では、共起ページ数、Jaccard 係数、overlap 係数、閾値付き overlap 係数のそれぞれを用いて共起の強さを計算した場合について、関係の強さ $R(X, Y)$ を横軸に、共著関係にある確率を縦軸にとったグラフを描いて評価・考察を行い、研究者の協働関係を推定する尺度として閾値付き overlap 係数が最も適していることを示している。以下、本論文で「overlap 係数」と書いた場合には、「閾値 $k = 30$ とした場合の閾値付き overlap 係数」のことを指す。さらに、[松尾 05] では、2 人の氏名でヒットした Web ページのテキスト解析を行うことで、共著、同研究室、同プロジェクト、同発表などの関係の種類を特定している。

Mika らは、この手法とほぼ同じ手法を提案している [Mika 04]。共起の強さを計るために Jaccard 係数 $R(X, Y) = |D_{X \cap Y}| / |D_{X \cup Y}|$ (ただし、 $|D_{X \cup Y}|$ は X, Y のいずれかを含む Web ページ数) を用いている点、テキスト解析による関係の種類を判別を行っていないなどの点で異なるが、基本的には同じアルゴリズムである。検索エンジンにより 2 つの語の間の関係性の強さを求める方法は、最近では比較的知られてきており、例えば Google Hacks [Calishain 03] の中で紹介されている。これらは、あらかじめ名前

*1 <http://www.google.com/apis/>

を与えたときに、その関連の強さを計る方法を基本としている。

これらの研究の起源となったのは、Kautz らの Referral Web の研究である [Kautz 97]。Referral Web では、まずシステムに入力した名前をクエリーとして検索し、その検索結果のページから人名を抽出し、関係の強さを Jaccard 係数で計算し、さらにそれらの人名をクエリーとして検索することを繰り返す。それにより、最初の名前を中心とした人のネットワークが抽出されることになる。また、原田らは、研究トピックや分野を表す検索語を入力とし、検索にヒットした最大 1000 件の Web 文書から人名を抽出し、その共起関係を調べている [原田 03]。[原田 03] では、検索語に関連する人物の関係を可視化することが目的であるため、人物の関係を調べる際に Web 全体を対象とせず、検索語に関連する最大 1000 件の文書における共起のみを調べている。これらは、Web の検索結果の（上位）ページから名前を抽出した上で、氏名の関係の強さを計算する方法である。

ここで、松尾らや Mika ら、前者の手法の問題となるのが、必然的にノード数 n に対して、 $O(n^2)$ の検索クエリーを検索エンジンに投げる必要がある点である。Referral Web では、この点が問題となり、最初に与えた氏名から数ノード以上のネットワークを抽出することは、ノード数が爆発してしまい、実質不可能である。原田らの方法は、あらかじめ対象とする Web 文書に 1000 件という上限を設定することで、この問題を回避している。

対象となる文書を限定することは、検索エンジンの負荷を下げることはできるが、関係が強いことを示す Web ページが対象文書以外に存在する可能性がある。例えば、学生同士が一緒に書いた授業のレポートを載せた Web ページは、その学生同士の関係性を示す証拠ではあるが、検索結果の上位に現れない可能性もある。一般的には検索エンジンの負荷を下げることで、抽出すべき関係を取り逃がさないことはトレードオフの関係にあり、なるべく検索エンジンの負荷を減らしながら取り逃がす関係を少なくする手法が望ましい。

3. 提案手法

本節では、Web 上の情報を用いて研究者ネットワークを抽出する際、検索エンジンの負荷と検索精度のトレードオフを解消する新しい手法を提案する。

3.1 提案手法の考え方

松尾らや Mika らの従来手法では、与えた氏名のリスト中の全ての組み合わせに対して、検索エンジンにクエリーを投げ、氏名が共起する Web ページの数を調べる。しかし、容易に想像がつくように、得られる関係性のネットワークは非常にスパースである。例えば、人工知能学会の 266 人のネットワークではエッジ数が 690 であり [松

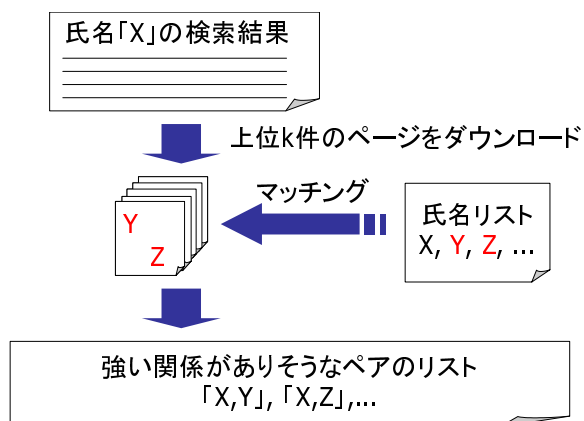


図 1 提案手法

尾 05]、ネットワークの密度（存在するエッジ数を可能なエッジ数で割ったもの）は 0.0195 と小さい。つまり、任意の 2 人のうち 2% 程度の間には強い関係は存在しておらず、すべての 2 人の組み合わせについて検索エンジンを利用するのは効率が悪い。

ひとつの方法として、関係が強い人を共有する（つまり共通の知人を持つ）2 者間にだけ、関係が存在するかどうかを検索エンジンを用いて調べるアプローチが考えられる。つまり、存在するエッジは強い紐帯 [安田 97] であると仮定し、ネットワークを徐々に広げていく方法である。しかし、人の関係で面白いのは、意外な人のつながりであり、すでに分かっている関係から存在するであろう関係を予測する方法は、ネットワークを把握する上で重要な「弱い紐帯」を取り逃がすことになる [安田 97]。

そこで、本手法では次のようなアプローチを取る。ある人を考えたときにその人と関係が強い人は、その人の氏名で検索した上位の文書におそらく含まれているだろうと考えることができる。例えば「浅田洋平」という氏名で検索したとき、その上位には、浅田洋平の所属する研究室のメンバーリストや発表文献、学会プログラムなどが含まれるが、浅田洋平と強い関係にある指導教員や共著者といった氏名は上位文書の中に含まれる。もちろん、偶然に同じ研究会で発表した人の氏名など、強い関係にない名前も含まれることになるが、少なくとも、強い関係にある人を取り逃がすことは少ないので、フィルタリングとして用いることができると考えられる。

この方法の概要を図 1 に示す。ある氏名「X」をクエリーとして検索された上位 k 件のページを取得し、氏名リスト中の氏名が含まれていないかを調べる。本手法では、氏名リストは与えられているので、氏名をパターンとする文字列マッチングを行うだけでよい。ただし、テキスト中では姓と名が半角スペースや全角スペース、タブなどで区切られていることがあるので、あらかじめテキストから半角スペース、全角スペース、タブなどを除去した上で処理を行う。そして、この上位 k 件のページに含まれる氏名（「Y」「Z」）との間には強い関係がありそうで

表 1 情報系研究者ネットワークのノードと抽出に必要なクエリ数

| 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 |
| 従来手法によるクエリ数 | 4142881 |
| 提案手法によるクエリ数 | 137967 |

あると推測し、これらの氏名との共起(「 X, Y 」「 X, Z 」)のみを検索エンジンで調べることにする。これにより、強い関係を取りこぼすことなく、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。

3.2 提案手法の流れ

提案手法は、次のようなステップを踏む。

Step1 ネットワークを構成するメンバーの氏名と所属のリストを用意する。

Step2 氏名リスト中の全ての氏名「 X 」をクエリとし、Web 検索を行い、その氏名を含む Web ページ数 $|D_X|$ を調べる。本研究では、[松尾 05] にならない、同姓同名の別人を除去するために「氏名 AND 所属」で検索した。このとき、検索結果における上位 k 件のページを解析し、氏名リスト中の氏名「 Y 」があれば、そのペア X, Y は共起ページ数を検索するペアのリストに追加する。

Step3 共起ページ数を検索するペアのリスト中の全てのペア X, Y について「 X AND Y 」をクエリとして Web 検索を行い、二人の氏名が共起する Web ページ数 $|D_{X \cap Y}|$ を調べる。検索していないペアについては、 $|D_{X \cap Y}| = 0$ とする。

Step4 リスト中の全ての氏名のペア X, Y について、それらの関係の強さ $R(X, Y)$ を overlap 係数 $R(X, Y) = |D_{X \cap Y}| / \min(|D_X|, |D_Y|)$ で計算する。

4. 大規模ネットワークの抽出

提案手法を用いて、実際に 3000 人規模のネットワークを抽出した。情報系の研究者を対象とし、表 1 に示すネットワークを抽出した*2。

抽出された大規模ネットワークの一部を図 2 に示す。ネットワークの表示には、Graphviz *3を用い、ばねモデルにより、ノード間の距離が overlap 係数の逆数を反映するようなレイアウトを求めている。図 2 では、955 のノードと、1936 本のエッジを表示している。ただし、2879 人全ての関係を抽出しているが、ネットワークを表示する際には、まず overlap 係数が 0.5 以上のエッジを実線で表示し、エッジ数が 3 本以下のノードについては、overlap 係数が 0.3 から 0.5 のエッジも破線で表示して

*2 2004 年 2 月時点での Google による結果。

*3 <http://www.research.att.com/sw/tools/graphviz/>

いる。この結果、孤立ノードになってしまった研究者は表示していない。多くの研究者と協働関係にある研究者の周りにはエッジが集中しており、黒くなっているのが分かるが、ネットワークの密度は 0.004 であり、任意の二人のうち、0.4%程度の間には強い関係はないことが分かる。ネットワークを抽出する際に要したクエリ数は表 1 の通りである。従来手法に対し、提案手法では、約 97%のクエリを削減できている。従来手法では、4142881 個のクエリが必要となるので、このネットワークを抽出することは困難である。

5. 評価実験

5.1 実験方法

提案手法は、氏名单独の検索結果の上位 20 件のページ中で共起している氏名とのペアについてのみ、強い関係がありそうなペアであるとして、検索エンジンに対する負荷を減らそうとするものである。すなわち、従来手法では全ての関係をもれなく抽出できるが、提案手法では共起ページ数を検索するペアのフィルタリングを行っているので、関係を取りこぼす可能性がある。したがって、提案手法を用いることで、従来手法に対して、

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

の 2 点に対して評価を行う。

前者の点については、共起ページ数の検索に要したクエリ数によって評価できる。

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氏名との共起ページ数しか調べていないので、弱い関係を抽出することを犠牲にしている。そこで、overlap 係数に閾値を設け、閾値以下の関係のないものとみなした場合に、何%の関係を抽出できたかを測る指標として、式 (2) のように被覆率を定義した。

$$\text{被覆率} = \frac{\text{提案手法による閾値以上の関係の数}}{\text{従来手法による閾値以上の関係の数}} \quad (2)$$

閾値を変えて被覆率を調べることで、提案手法がどの程度関係を何%抽出できたのかということの評価できる。

本評価実験では、500 人程度の規模のネットワークを従来手法、提案手法の二つの方法で抽出し、結果を比較した。検索エンジンとしては代表的な検索エンジンの一つである Google を用いた。実験で抽出したネットワークのノードを表 2 に示す。提案手法では、 $k = 20$ 、すなわち、氏名单独の検索結果の上位 20 件のページに出現する氏名との共起ページ数についてのみ検索を行った。

また、本評価実験では、overlap 係数を計算する際に、閾値付き overlap 係数を用いたため、従来手法、提案手法のいずれの場合も、ヒット件数が 30 件以下の研究者との関係については抽出していない。閾値付き overlap 係数と、これを用いる理由については、2 節 (4) を参照されたい。

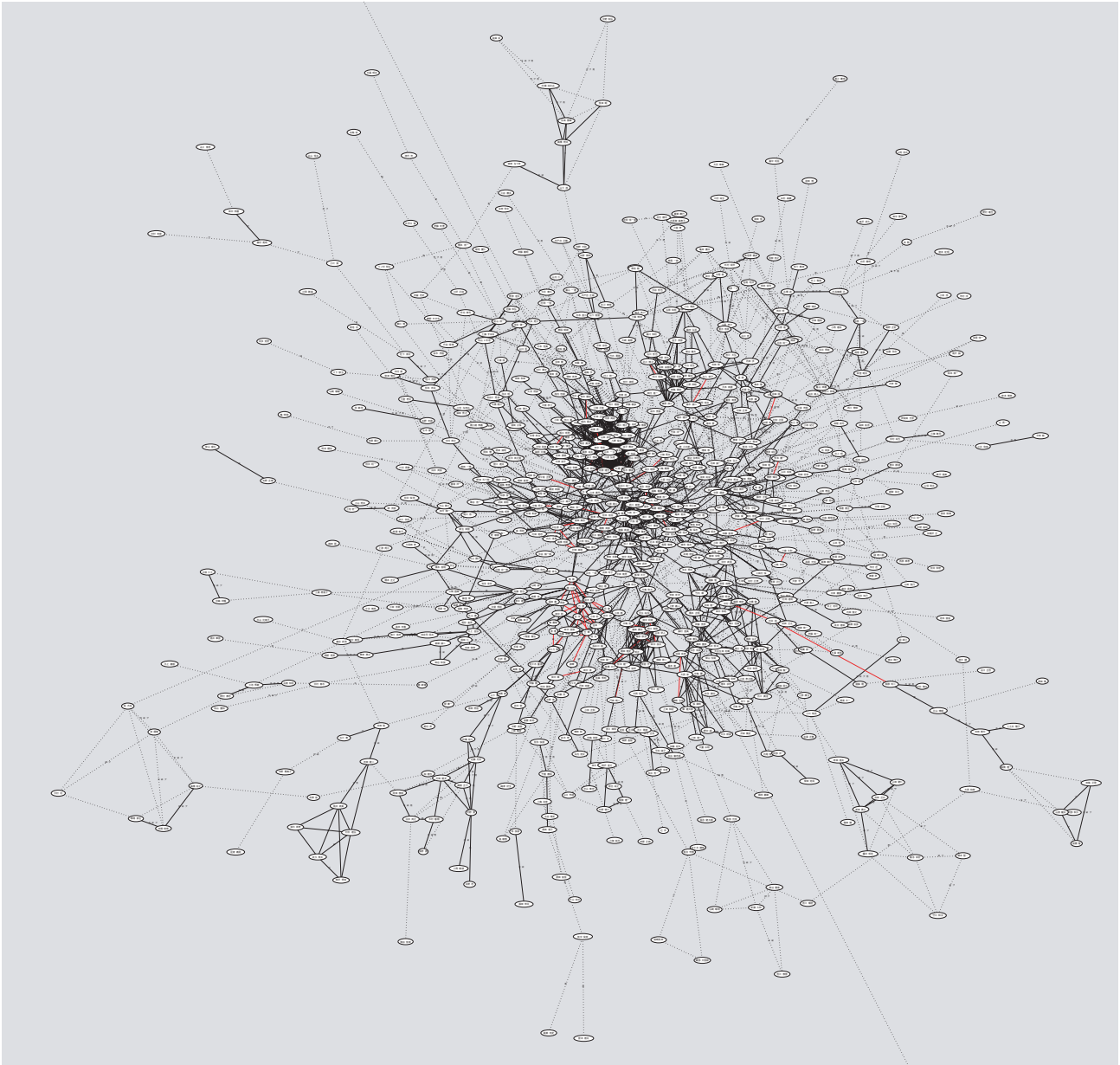


図 2 情報系研究者のネットワーク（一部）

表 2 実験で抽出したネットワークのノードとネットワーク抽出に要したクエリ数

| ノード | JSAI2003 の参加者 |
|-------------|---------------|
| ノード数 | 503 |
| 従来手法によるクエリ数 | 126253 |
| 提案手法によるクエリ数 | 19182 |

5・2 結 果

§1 提案手法 ($k = 20$) の評価

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検索エンジンに与えたクエリ数は表 2 に示す通りである。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。また、表 1 と表 2 を比べると、従来手法で 500 人程

度のネットワークを抽出するのと同程度のクエリ数で、提案手法では 3000 人程度のネットワークを抽出できることが分かる。

次に、提案手法により抽出できた関係の強さの傾向を調べる目的で、全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、横軸を従来手法による overlap 係数、縦軸を提案手法による overlap 係数とするグラフ上にプロットした。結果を図 3 に示す。プロットされた点は、任意の二人の研究者の関係を意味している。図 3 では、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、横軸上のドットが、提案手法によって抽出できなかった関係を示している。図中で、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

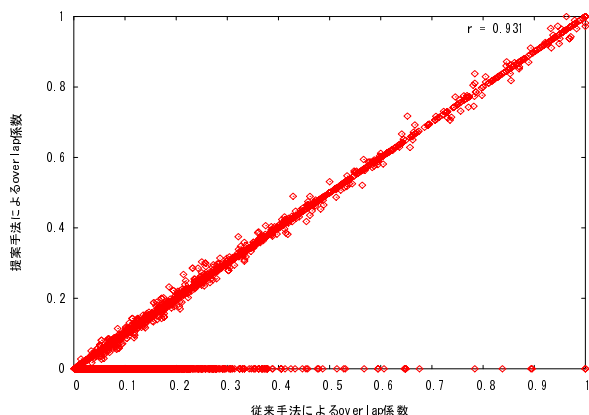


図 3 従来手法と提案手法の相関

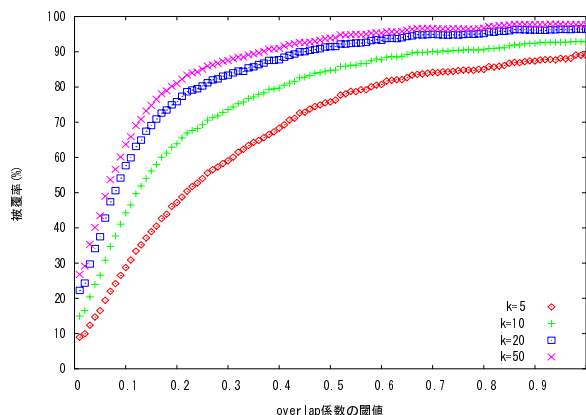
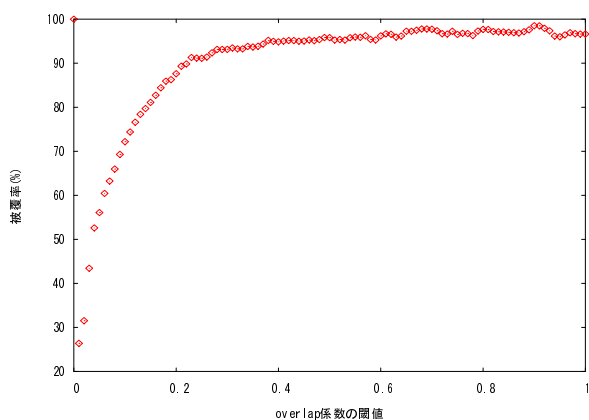
図 5 k を変化させたときの被覆率の変化

図 4 閾値による被覆率の変化

ることによるものである。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、提案手法での取りこぼしが少ないことが分かる。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、overlap 係数の閾値を 0 から 1 まで 0.01 刻みに変化させた場合の被覆率を式 (2) によって求めた。結果を図 4 に示す。図 4 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% が抽出できていることが分かる。[松尾 05] では、人間関係ネットワークを表示する際に、まず overlap 係数が 0.7 以上のエッジを表示し、総エッジ数が 3 本以下のノードについては overlap 係数が 0.5 から 0.7 までのエッジも表示するようにしている。対象とするコミュニティの規模や種類によって、表示に適切な overlap 係数の閾値は異なると考えられるが、例えば、提案手法で 0.2 以上のエッジを全て表示した場合、従来手法の約 88% のエッジを表示できることになる。

以上をまとめると、提案手法を用いると、従来手法に比べて

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

という評価ができる。すなわち、提案手法は、強い関係の取りこぼしを最小限に抑えつつ、検索エンジンに対する負荷を大幅に減少できていることが分かる。

§2 パラメータに関する評価

k の値を変化させたときに被覆率がどのように変化するかを調べた。この実験では、検索エンジンのデータベース更新により共起ページ数が変化し、抽出された関係の overlap 係数が手法によって異なってしまう問題を避けるため、提案手法の overlap 係数は、共起ページ数を検索したペアについては従来手法と同じ値、それ以外の場合には 0 とした。 $k = 5, 10, 20, 50$ の 4 通りについて被覆率のグラフを描くと、図 5 のようになった。 k を大きくするほど被覆率が改善されている、すなわち、多くのページを解析するほど多くの関係を抽出できていることが分かる。また、 $k = 5$ と $k = 10$ 、 $k = 10$ と $k = 20$ の系列の差に比べて、 $k = 20$ と $k = 50$ の系列の差は小さくなっている。このことは、強い関係にある人の氏名が検索結果上位のページに多く出現し、ある程度以上に k を大きくしても被覆率の改善は小さいことを示していると考えられる。 k を大きくすると、上位 k ページ中でより多くの氏名と共起することになり、氏名の共起ページ数を検索するのに必要なクエリ数も多くなるので、抽出するネットワークの種類や規模に応じて k の値を決めればよいことが分かる。

次に、ネットワークのノード数 n に対する、共起ページ数の検索に要するクエリ数のオーダについて理論的に考察する。

氏名单独をクエリとして検索された上位 k 件の Web ページ中で共起している氏名リスト中の氏数の数が一人当たり平均 m 人 ($m \leq n$) であるとする。共起ページ数の検索に要するクエリ数は、高々 $n \times m$ である。

ノード数 n の増加にしたがって、共起氏数名 m も増加するので、明らかに m は n の関数である。 m の関数を推定することは難しいが、 $k = 20$ とした場合、JSAI2003 参加者のネットワークでは $m = \frac{19182}{503} \approx 38$ 、情報系研究者のネットワークでは $m = \frac{137967}{2879} \approx 48$ である。この例

を見る限り、 n の増えかた (5.7 倍) に対して m の増え方 (1.3 倍) は緩やかであるし、一つの氏名と共に起る氏名の数には上限があるため、提案手法によるクエリ数は $n \geq 3000$ のネットワークの抽出に対しても、ほぼ $O(n)$ の増加に抑えられると考えられる。

6. 議 論

ソーシャルネットワークの抽出について、これまで様々な研究が行われてきた。

Web を情報源としてソーシャルネットワークを自動的に抽出する研究としては、松尾らの研究 [松尾 05, 友部 04] の他にも、Referral Web [Kautz 97] や、Mika らの研究 [Mika 04]、原田らの研究 [原田 03]、SocialPathFinder [Ogata 99] などがある。本節では、それらの関係をまとめ関連付けるために、やや詳しく各研究を紹介する。

Referral Web [Kautz 97] は、ソーシャルネットワークを用いた情報検索、エキスパート検索システムである。ユーザがシステムに氏名を登録すると、システムは検索エンジン^{*4}を用いてその氏名を含む Web ページを検索する。次に、検索された Web 文書から固有表現抽出により、氏名を抽出し、氏名リストを作成する。そして、ユーザが入力した氏名「 X 」と、氏名リスト中の氏名「 Y 」の全てのペアについて「 X AND Y 」、「 X OR Y 」をクエリとした Web 検索を行い、検索されたページ数をそれぞれ $|D_{X \cap Y}|$ 、 $|D_{X \cup Y}|$ とすると、関係の強さを Jaccard 係数 ($\frac{|D_{X \cap Y}|}{|D_{X \cup Y}|}$) で計算する。さらに氏名リスト中の氏名をシステムに入力する、ということを i 回繰り返すことで、ユーザを中心とする半径 i のネットワークを得ることができる。このネットワークを蓄積していくことで、大規模なネットワークを抽出する。しかし、この手法では、あらかじめ氏名リストを用意せず、固有表現抽出により氏名リストを抽出するため、例えば情報系の研究者ネットワークを抽出したい場合にも、情報系ではない研究者および、研究者でない人もネットワークに含まれてしまうことになる。また、全体のネットワークは、各ユーザの半径 i のネットワークを統合したもので、氏名をクエリとした検索を行っていないノードについては、重要な関係を取りこぼしている可能性があり、網羅的ではない。検索回数については、一つの氏名と共に起る全ての氏名の数が平均 x であると仮定すると、ノード数 n のネットワークを抽出するためには高々 $n \times x$ の検索クエリが必要となる。 x は非常に大きな数になると考えられるが n に依存しない定数であるため、検索回数は提案手法と同じ $O(n)$ である。ところで、 x が一つの氏名と共に起る全ての氏名の数であるのに対して、 m はそのうち上位 k 件のページ中で共起する、氏名リストに含まれる氏名の数であるから、明らかに $x > m$ である。すなわち、提案

手法の検索回数は 5.2 節で述べたとおり高々 $n \times m$ であるから、Referral Web の検索回数 $n \times x$ は提案手法よりも多い。

Mika らは、従来手法とほぼ同様の手法を提案している [Mika 04]。すなわち、あらかじめ氏名リストを用意し、検索エンジンで氏名の共起ページ数を調べる。共起の強さには Jaccard 係数を用いている点、関係にラベルをつけていない点が松尾らの手法と異なるが、基本的なアルゴリズムは同様である。したがって、従来手法と同様、ネットワークのノード数 n に対して $n C_2$ の検索クエリが必要となり、大規模なネットワークの抽出が困難であるという問題がある。

原田らは、あるトピックに関するキーパーソンの関係を抽出するシステムを提案している [原田 03]。ユーザがトピックを入力すると、システムは検索エンジンでそのトピックをクエリとする Web 検索を行い、最大で上位 1000 件の Web ページを取得する。次に、これらのページから固有表現抽出によって人名を抽出し、それぞれの人名について、検索結果の Web ページの集合とその人名を含む全ての Web ページの集合の重なり度合いを対数尤度比検定で用いられる G スコア [Dunning 93] で計算し、これをその人名とトピックの関連度とする。さらに、Web ページ中での共起の距離に基づいて人名間の関係の強さを計算し、トピックに関連する人間の間関係を抽出している。この手法では、人名間の関係の強さを計算する際に検索エンジンを用いないため、クエリ数については問題がない。しかし、対象とする情報源をトピックに関連する最大 1000 件の Web ページに限定しているので、重要な人物や重要な関係を取りこぼす可能性がある。すなわち、ノード間関係は網羅的ではない。

SocialPathFinder [Ogata 99] は、ユーザの Web ページに存在するハイパーリンクによって人間関係を抽出するシステムである。システムはユーザのホームページを起点として、ハイパーリンクをたどり、リンク先のページが個人のホームページであるかどうかを URL やメールアドレスの情報から推定する。個人のホームページにおけるリンク関係を用いることで、プライベートな関係を抽出することができるが、ホームページを持っていない人との関係は抽出できない。また、個人の名前は URL やメールアドレスから抽出するので、「ogata」「okawa」といった本人以外には分かりにくいものであるだけでなく、そもそも URL やメールアドレスが個人名になっていない場合には関係を抽出できない。すなわち、重要な人物、関係を取りこぼす可能性があり、ノード間関係は網羅的ではない。

以上のことから、提案手法も含めてそれぞれの手法の相違を表にまとめると、表 3 のようになる。表 3 から、関係の網羅性が低いと、検索エンジンの負荷が小さくなることが分かる。すなわち、ネットワークを取りこぼしなく網羅的に抽出することと、検索エンジンに対する負荷

*4 AltaVista(<http://www.altavista.com/>) を用いている。

表 3 Web を用いたソーシャルネットワーク抽出手法の比較

| | 松尾らの手法 | Referral Web | Mika らの手法 | 原田らの手法 | SocialPathFinder | 提案手法 |
|------------|------------|--------------|------------|-------------|----------------------|------------|
| 関係の根拠 | 氏名の共起 | 氏名の共起 | 氏名の共起 | 氏名の共起 | ハイパーリンク | 氏名の共起 |
| 関係の強さ | overlap 係数 | Jaccard 係数 | Jaccard 係数 | 共起距離 | なし | overlap 係数 |
| 氏名リスト | あらかじめ用意 | 固有表現抽出 | あらかじめ用意 | 固有表現抽出 | URL, E-mail アドレスから抽出 | あらかじめ用意 |
| 情報源 | Web 全体 | Web 全体 | Web 全体 | 最大 1000 ページ | 個人の Web ページ | Web 全体 |
| 検索エンジンへの負荷 | $O(n^2)$ | $O(n)$ | $O(n^2)$ | 1 | なし | $O(n)$ |
| 関係の網羅性 | | x | | x | x | |

を減らすことは一般的にトレードオフの関係にあり、従来の研究は、

- 情報源を絞りこむことで関係の網羅性を犠牲にするもの: 原田らの手法, SocialPathFinder, Referral Web (情報源は Web 全体としているが, 重要な関係を取りこぼす可能性があるため, こちらに分類した)
- 関係の網羅性を優先することでクエリ数が多くなるもの: 松尾らの手法, Mika らの手法

の二つに分けられることになる。

本論文の提案手法は, 情報源は Web 全体としながらも, 検索エンジンに投げるクエリに対してフィルタリングを行うことで, 網羅性を保ちつつ検索エンジンへの負荷を減らすシンプルかつ有効なアプローチであると考えられる。

また, ソーシャルネットワーク抽出以外の分野でも, Web マイニングの研究では, 検索エンジンを用いて語の共起を調べる研究が行われている。例えば, Tuney は, 語と語の意味的な関係の強さを Web 上での共起にもとづいて推定する手法を提案している [Tuney 03]。しかし, Tuney 自身が言及しているように, この場合もソーシャルネットワークの抽出と同様に検索エンジンへのクエリ数が問題となっている。

このように, Web から検索エンジンを用いて関係を抽出する手法は有効なものであるが, 潜在的にクエリ数の問題が存在しており, 今後, Web マイニングの分野において, 本論文の提案手法のようなアプローチの必要性が高まると考えられる。

7. おわりに

Web 上の情報からの研究者ネットワークの抽出手法の従来手法において, 検索エンジンに与えるクエリをフィルタリングすることでクエリ数を減らし, 大規模なネットワークの抽出を可能にする手法を提案した。また, JSAI2003 の参加者を例にとって従来手法と提案手法を用いてネットワークを抽出し評価を行い, 提案手法により, 強い関係の取りこぼしを最小限に抑えつつ, 検索エンジンに対する負荷を従来手法に対して大幅に削減できることを示した。本論文では研究者を対象として Web 上での共起を調べているが, 提案手法は, Web における名前の共起を調べる際に, 広い適用範囲に対して有効な可

能性があると考えている。

謝 辞

本研究を行う上でお世話になった全ての方に心より感謝いたします。

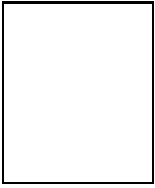
◇ 参 考 文 献 ◇

- [Calishain 03] Calishain, T. and Dornfest, R.: *Google Hacks*, O'reilly (2003)
- [Dunning 93] Dunning, T. E.: Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics*, Vol. 19, No. 1, pp. 61-74 (1993)
- [藤井 04] 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)
- [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, 情報処理学会研究報告 DBS-130/FI-71 (2003)
- [Kautz 97] Kautz, H., Selman, B., and Shah, M.: The hidden web, *AI Magazine*, Vol. 18, No. 2, pp. 27-36 (1997)
- [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. 1, pp. 46-56 (2005)
- [Mika 04] Mika, P.: Bootstrapping the FOAF-Web: an experiment in social network mining, in *Proc. 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web* (http://www.w3.org/2001/sw/Europe/events/foaf-galway/papers/fp/bootstrapping_the_foaf_web/) (2004)
- [Mori 04] Mori, J., Matsuo, Y., Ishizuka, M., and Faltings, B.: Keyword extraction from the Web for person metadata, in *Poster Abstracts 3rd Int'l Semantic Web Conference (ISWC2004)*, pp. 45-46 (2004)
- [Ogata 99] Ogata, H., Fukui, T., and Yano, Y.: Social-PathFinder: computer supported exploration of social networks on WWW, in *ICCE99, vol.2*, pp. 768-771 (1999)
- [友部 04] 友部 博教: 知識共有システムにおける知識の獲得・加工・管理に関する研究, PhD thesis, 東京大学大学院情報理工学系研究科 (2004)
- [Tuney 03] Tuney, P. D.: Coherent keyphrase extraction via web mining, in *Proc. of the 18th Intl. Conf. on Artificial Intelligence (IJCAI-03)*, pp. 434-439 (2003)
- [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)
- [安田 04] 安田 雪: 人脈づくりの科学, 日本経済新聞社 (2004)
- [吉川 02] 吉川 弘之: 科学者の新しい役割, 岩波書店 (2002)

[担当委員: x x]

19YY 年 MM 月 DD 日 受理

—— 著 者 紹 介 ——



著者 1 姓 名(正会員)

著者 1 の略歴