語の共有に基づく文書ネットワークの構造的特徴について

佐藤進也[†] 福田健介[††]
風間一洋[†] 村上健一郎[†††]

自然界，社会に存在する様々なシステムや，Web のような，人工的でありながら個々の要素が自律的に振る舞っているシステムを記述・解析する方法として，要素間の相互関係をネットワークで表現し，その構造を解析するというアプローチが有望視されている．一般に，要素間には複数の関係が存在し，その中のどの関係に注目するかによって異なった構造のネットワークが得られ，それぞれから当然異なった解析結果が得られる．つまり，解析結果の質（正確さなど）は解析手法だけでなくネットワークの構成方法にも依存する．よって，意味のある結果を得るためには「よい」ネットワークを構成する必要があり，そのための方法論の確立が望まれる．そのためにはまず「よい」ネットワークの要件を明らかにする必要がある．本論文では，その要件をネットワークの構造的特徴として定義しうることを，コーパス中の文書の相互関係を表すネットワークを題材にして示す．

On the Structural Characteristics of Document Networks Based on Sharing of Terms

Shin-ya Sato,[†] Kensuke Fukuda,[††] Kazuhiro Kazama[†]
and Ken-ichiro Murakami[†††]

For describing and analyzing a system in nature, society as well as an artificial systems consisting of autonomous elements, it is considered to be a promising approach to analyze the structure of the network representing interrelationships among its elements. In general, there may be various types of relationships among elements, each of which gives a network with a peculiar structure, and a result of analysis based on the structure of the network. This means that the quality (e.g., reliability) of the result depends not only on the analytical method but also on the method to build the network. Therefore, we need to devise a method to construct a *good* network to obtain a valid result, which requires a measure for *goodness* of a network. Through an experiment with networks representing interrelationships among documents in a corpus, we show that such measures can be defined in terms of structural characteristics of the networks.

1. はじめに

生物や市場経済などの自然界，社会に存在する複雑な振舞いを示すシステムや，インターネットや Web などの人工的でありながら個々の要素が自律的に振る舞うシステムを，その要素間のつながり（ネットワーク）に着目して解析し理解しようとする動き，いわばネットワーク指向アプローチが近年注目されている．社会科学においては，以前より人どうし，組織どうしなどの関係をネットワークで表現し，その構造を解析するネットワーク分析というアプローチが用いられてきた[1]．また，Web 情報検索においても，情報（Web ページ）間のつながりに注目した研究が数多く行われている．Web ページの「重要度」をそこに張られているリンクの数をもとに計算するもの[2]，2 つのページの類似性をその他のページからのリンクの張られ具合いで推定するもの[3] などをその例としてあげることができる．これらの研究手法には，それぞれの要素の性質（特徴）を他の要素との相互関係により表現するという共通点があり，それがネットワーク指向アプローチの 1 つの重要なポイントであるといえる．さらに最近ではスモールワールド[4] やスケールフリー[5] といったネットワークの大域的性質が注目されており，その関連の研究がさかんに行われている．

† NTT 未来ねっと研究所
NTT Network Innovation Laboratories
†† 国立情報学研究所
National Institute of Informatics
††† 法政大学ビジネススクールイノベーション・マネジメント研究科
Hosei Business School of Innovation Management

さて，ネットワーク指向アプローチの解析対象は，ネットワークがあらかじめ与えられているものと，ネットワークの構成が解析にともなって必要となるものの 2 つに大別される．たとえば，先ほどの Web 情報検索の例では，Web ページの重要度を測ることとは独立に，Web ページを頂点としハイパーリンクを辺とするネットワークがすでに存在している．一方，人間関係の解析では，まずネットワークを構成するところから始めなければならない．人どうしの関係を示す要素は様々であり，どの要素に注目するかによってネットワーク構造も変化し，それは解析結果にも影響を及ぼす．よって，ネットワークをいかに構成するかは重要な問題であり，そのための何らかの指針が示されることが望まれる．

そこで，本論文では，ネットワーク指向アプローチの観点から見て「よい」ネットワークの構造的特徴を探る．ここでは，ある方法により対象（たとえば人間関係）をネットワークで表現し，それを解析して得られた結果が事実とよく合致するとき，それを「よい」ネットワーク（の構成方法）であると考える．「よい」ネットワークとそうでないネットワークを構造的特徴に基づいて区別できるか，という問いに答えることが本論文の目標である．

この議論のため，本論文では題材として文書のネットワークを考える．これは，それぞれの文書をそこに含まれる語によって特徴付けし，同じ語を共有するか否かによって相互の関係（すなわちネットワーク）を決定するものである．この場合，ネットワークの構造は特徴語の選び方に依存する（詳しくは 2 章参照）．我々は，「よい」ネットワークを作るための語の選び方についてはすでに知見を得ている[15]．この結果をふまえて，本論文では，特徴語（の選び方）とそれによって得られるネットワーク構造の関係を調べる．

以下，次のように議論をすすめていく．まず，2 章では，関連研究として，文書とそこに含まれる語の関係，そのネットワークによる表現と応用に関する既存のアプローチを紹介する．さらに，これらの先行研究では議論されることのなかった，ネットワークの構成方法の重要性について述べる．この問題提起をうけて，3 章では，ネットワークの構成方法を評価する基本的手順を示す．その手順の前半に対応する部分，すなわち，異なるタイプのネットワークの構成と，その妥当性の評価結果を 4 章において述べる．手順の後半部分は 5 章で議論する．ここでは，前章で得られたネットワークを比較することにより，「よい」ネットワークの構造的特徴を探る．そして，その特徴に基づいた取捨選択により，実際に，「よい」ネットワークが抽出できることを示す．

2. 関 連 研 究

2.1 語の共起と共有

一般に，1 つの文書で述べられていることには一貫性があるので，同じ文書内に現れる 2 つの語（が示す概念）には関連性があると考えられる（文書内における語の**共起**）．また，同様な議論から，同じ語を含む 2 つの文書にも関連性が期待される（文書による語の**共有**）．しかし，実際に語の共起や共有といった事実から語や文書どうしの関連性を導くためにはさらに工夫が必要である．たとえば，**図 1** の (A) に示すように，5 つの文書 V〜Z に語 a〜f が出現しているとする．ここで，語 a はすべての文書にあまねく出現しているが，この状況を，語 a は文書の内容にかかわらず出現しているととらえることができる．いい換えれば，a を含むという事実は文書の類似性を判断する材料にならないということである．このように，ある語の共起や共有から類似性を導くには，それが有意なものであるかを判定しなければならない．この有意性の判定については，従来より，Jaccard 係数[6]，G-score[7]，χ^2 検定法の応用[8] など，語の出現に関する統計量に基づいて基準を決める手法が数多く研究されてきた．

2.2 ネットワーク指向アプローチ

さらに，言葉の意味や文書に記述されている内容をより正確に把握するためには，文書や語の個々の関係ではなく複数の関係の総合的把握が必要であることが認識され，その手段として，ネットワーク指向アプローチの適用が試みられてきた．このアプローチで用いられるネットワークの代表例が，共起関係でつなげた語のネットワークと語の共有関係でつなげた文書の

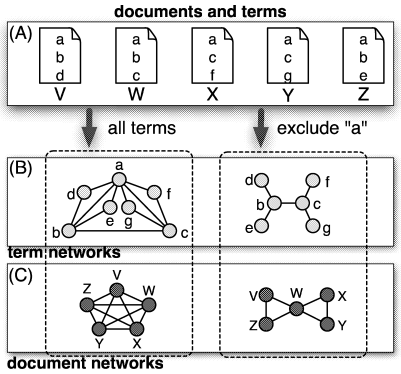

図 1 共起・共有関係を表す語と文書のネットワーク
Fig. 1 Networks of terms and documents representing cooccurrence and sharing of terms.

ネットワークである．図 1 のすべての語と文書の関係を使ってこれらのネットワークを構成すると，語のネットワークは (B)，文書のネットワークは (C) となる（それぞれ左側の破線枠内）．これらのネットワークは，まず，従来手法では認識できない関連性を見つけ出す手段として利用できる[9]．たとえば，図 1 で共起関係だけを調べると，語 d と語 e には関連性がないと判断されてしまう．一方，語のネットワーク上でこの二者の関係をみると，b を仲立ちとして間接的につながっていることが分かる．いい換えれば，d と b，e と b にそれぞれ関連性があることから d と e の間にも関連性が示唆される，ということである．もちろん，この間接的なつながりの有意性については別に議論が必要ではあるが，少なくとも，個々の共起関係に注目していただけでは分からなかった関連性（の可能性）を発見可能にした点で，このアプローチは有意義であるといえる．

さらに，これらのネットワークの構造は意味を表現する手段としても利用できる[10]．たとえば，多義性のある語は，語のネットワークを（局所的に）複数の部分に分解する分岐点として特徴付けられる．これは，それぞれの語義ごとでは関連する語が互いに結び付けられ稠密な構造を形成する一方で，異なる語義間に結び付きがまたがることは非常に稀であるからである．このように，ネットワークの構造には，意味の同異性，関連性が反映されている．この性質を利用して，情報（文書集合）の理解を支援するために出現語のネットワークを可視化するという手法が考えられている[11]．また，同一文書中ではなく，同一文中に範囲を絞って語の共起のネットワークを構成し，その構造的特徴から，文書の主題を示すような重要語を抽出するというアプローチもある[12),13)]．

2.3 ネットワーク構成方法の妥当性

このように，語の共起や共有に基づく語，文書のネットワークは有用な情報を含んでいる．この情報を引き出すために，既存研究では主にネットワークの解析手法を工夫してきた．しかし，ネットワークの構成方法も解析方法に劣らず重要な課題である．このことを，再度，図 1 の例で考えてみる．ここには，語と文書のネットワークがそれぞれ 2 種類示されている．左の破線枠内は，文書に出現する語をすべて用いてネットワークを構成した場合であり，右の破線枠内は，語 a 以外でネットワークを構成した場合である．この 2 種類のネットワークには構造的に大きな違いがあるのが一見して分かる．このように，語の取捨選択はネットワークの構造，ひいてはその解析結果に少なからぬ影響を与えるのだが，その方法についていまだネットワーク構造の観点から議論されることはなく☆，文書中の語を TF・IDF☆☆などの統計的指標に基づいてランキングし，その上位を選ぶという標準的な特徴語抽出手法[14)] が流用されてきた．

語の選択基準は統計的指標以外にも考えられるため☆☆☆，統計量に基づく選択方法を採用することも，さらに，その方法に従って 1 文書あたりどのくらいの個数を選び出すかも恣意的であり，その妥当性を何らかの方法で客観的に示すことが望まれる．この要求に応えることを目指し，本論文では，語の選び方，すなわちネットワークの作り方の妥当性をネットワークの構造的特徴としてとらえる可能性を探る．

3. 手　　順

本論文では，以下の手順により，ネットワークの作り方の妥当性と構造的特徴との関係を調べる．

(1) 課題の設定

ネットワーク指向アプローチを用いて解決すべき課題を 1 つ設定する．具体的には，本論文では，「**文書ネットワーク**を用いた **同姓同名人物の分離**」という課題を考える．前章で，多義性のある語がネットワークを複数の部分に分解する分岐点として特徴付けられることを述べたが，同姓同名人物の名前もまた同様な性質を持つ語としてとらえることができる．つまり，ある人名が出現する文書の集合に対して適切な関係を導入することで，構造的に分解可能でそれぞの部分が個々人に対応しているようなネットワークを構成できると考えられる．なお，ここでは語のネットワークではなく，文書のネットワークに注目しているので，人名ごとにネットワークの頂点は固定され，構成方法の違いは頂点間の関係，すなわち辺の張られ方の変化に対応する．

(2) ネットワーク構成方法の妥当性の検証

複数の方法でネットワークを構成し，それぞれを用いて (1) の課題を解く．その結果がどの程度正しいかを

☆ 先にも述べたように，1 つの文書中に出現する語の共起関係をネットワークで表現し，その構造に基づき重要な語を選択するという先行研究はあるが，いま議論しているのは，他の文書との関係をネットワークとして記述するのに適した語をどのように選ぶかという問題である．

☆☆ TF（term frequency）とは 1 つの語の 1 文書中での出現頻度のことであり，IDF（inverse document frequency）は全文書における当該語の出現の “稀さ” を示す数量で $\log N/DF$ などといった式により表される．ここで，N は全文書数，DF は当該語が出現する文書の数である．

☆☆☆ たとえば，固有名詞などといった語の（形態素としての）種別に基づくもの．

事実と照らし合わせて判断する．いま我々が解くべき問題は同姓同名人物の分離であり，文書のネットワークを部分に分解し，それぞれが各個人に対応する（その部分に属する文書に記されている当該人名が同一の人物を指し示す）ようにすることである．この課題に対して，より正確な結果を導くネットワークが「よい」ネットワークであり，そのようなネットワークを構成する方法をより妥当であると判断する．

(3) 妥当性とネットワーク構造との対応付け

手順 (2) で構成したそれぞれのネットワークの構造的特徴を調べ，構成方法の妥当性との関連性を検討する．

手順 (1)，(2) については先行研究[15),16)] においてすでに結果が得られており，その概要を次章で紹介する．手順 (3) については，5 章で議論する．

4. 文書ネットワークの構造的特徴を利用した同姓同名人物の分離

4.1 ネットワークの構成方法

ネットワーク構成方法の基本方針を改めて端的に述べると次のようになる．すなわち，人名 x が与えられたとき，x を含む文書を頂点とし，それらの文書間で（x 以外の）語が共有されるとき辺を生成するというものである．この詳細を書き下すと次のようになる．

まず，文書の集合 D_0 を用意する．また，D_0 に属する文書で語 x を含むものの集合を $D(x)$ とする．$D(x)$ に含まれる各文書 d から

($\mathcal{G}$) **一般語**（general term）：形態素解析により，名詞あるいは未定義語と判断されたもの，

($\mathcal{P}$) **人名**（personal name）：形態素解析により姓，名と判断された語が連続して出現するとき，この 2 つをまとめて人名とする，

のいずれかを選び出し，異なる構造を持ったネットワークを構成する．具体的には，d ごとに☆一般的な語（あるいは人名）を取り出して TF・IDF 法で順位付けし，その上位から n_{max} 個を選びだす．これを T_d とする．T_{d_1} と T_{d_2} が要素を共有するとき，2 つの文書 d_1 と d_2 を辺で結ぶことにより，ネットワークが得られる．文書集合 D_0 を固定したとき，このネットワークは人名 x，語の種類，そして語数の上限に依存するので，

$G_{\mathcal{G}}(x, n_{max})$ あるいは $G_{\mathcal{P}}(x, n_{max})$

と書くことにする．ただし，n_{max} の値は，実際にどのくらいの数の語が用いられているかを正確に示しているわけではないことに注意しなければならない．たとえば，出現する語の数がいずれの文書でもたかだか 10 である状況では，$G_{\mathcal{G}}(x, 10)$ も $G_{\mathcal{G}}(x, 100)$ も同じネットワークを表すことになる．そこで，語数に関する状況を示したい場合には，文書あたりの語数の平均

$$\overline{n} = \frac{1}{|D(x)|} \sum_{d \in D(x)} |T_d|$$

を付加して

$G_{\mathcal{G},\overline{n}}(x, n_{max})$

あるいは n_{max} を省略して

$G_{\mathcal{G},\overline{n}}(x)$

などという表記を用いることにする．

4.2 ネットワーク解析手法

さて，n_{max} の値を大きくとると辺の数の増加が見込まれるが，一方，頂点の数は一定なので，ネットワークは込み合ってくると想像される．**図 2** はある $G_{\mathcal{P}}(x, n_{max})$ の実際の変化を示したものであるが（(i) から (iv) に向けて n_{max} が増えている），その予想が正しいことが見てとれる．

これらのネットワーク構造で注目すべきなのは，(ii) と (iii) に明らかに認められる，辺が密集している 2 つの部分である．このような，内部に存在する辺が外部とのつながりを与える辺よりも密に張られているような頂点の集合はコミュニティと呼ばれる[17)]．解析の結果，実は，これらのコミュニティはそれぞれ異なる 2 人の人物に対応していることが分かっている．

このことから，同姓同名人物分離はネットワーク中のコミュニティの存在を認識し抽出することに帰着できると考えられる．そのために着目すべきネットワークの構造として，**連結性**と**稠密性**があげられる．

最も単純な方法は，連結しているものをひとまとまりにすること，すなわち，ネットワークを各連結成分ごとに分解するというもので，(ii) のような状況でそ

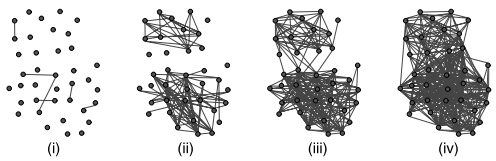

図 2 頂点間の関係の増加にともなうネットワーク構造の変化
Fig. 2 Transformation of a network with increase in number of relationships.

☆ 厳密には，文献 15)，16) では文書として Web ページを用いており，Web サイト上で（ファイルとして）近接しているページどうしはグループ化し，1 つのページと見なして処理している．近接するページ群は，共通した目的のもとに作成されていることが多く，それゆえ，高い関連性を持つという経験則に基づいている．また，ネットワーク解析の観点からいえば，これは，近接ページ間の “自明な” 関係をあらかじめ取り除き，粗視化するという意味を持つ．

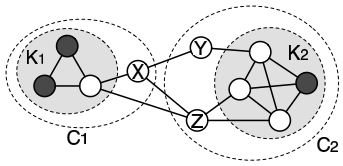

図 3 稠密性に注目したネットワークの分割
Fig. 3 Dividing a network based on density.

の有効性が期待できる．

一方，(iii) のような，つながりが増えて全体が連結している状況では，さらに詳細な解析が必要になる．そこで，稠密性に着目した次の方法を考える．まず，頂点どうしが最も稠密な関係を持っている部分，具体的にはクラスタ係数[4] が 1 である頂点とその近傍，を見つけ出し，コミュニティの核とする．そして，その他の部分は（複数ある）核のうち最も近いものに帰属させる．**図 3** のネットワークを例にとると，塗り潰した 3 つの丸印がクラスタ係数 1 のノードであり，それら近傍 K_1，K_2 をコミュニティの核とする．残りの頂点 X，Y，Z は，K_1，K_2 のいずれか近い方に帰属させる．このとき，つながりの強さも考慮する．たとえば，頂点 Z は K_1，K_2 から等距離にあるが，K_1 とは 1 つ，K_2 とは 2 つの辺でつながっているため K_2 とのつながりがより強いと考える．かくして，このネットワークは C_1 と C_2 という 2 つのコミュニティに分解される．

同姓同名人物分離の解法は，グラフの構成方法（あるいはそれにより構成されるグラフ）とネットワーク解析手法の組合せで表すことができる．連結性（connectivity），稠密性（density）に着目した解析手法をそれぞれ $\mathcal{C}$，$\mathcal{D}$ で表すことにすると，たとえば，人名 x に関する一般語を用いたネットワークの連結性に着目した解析 $\mathcal{A}$ は

$$\mathcal{A} = \{G_{\mathcal{G}}(x, n_{max}),\ \mathcal{C}\}$$

と書ける．個々の解析ではなく，任意の人名に対する解法としてとらえた場合には，これを

$$\mathcal{A} = \{\mathcal{G}, \mathcal{C}\}$$

と抽象化して書くことにする．

4.3 評　　価

4.3.1 認　識　率

解法の妥当性は，ネットワークから抽出したコミュニティがいかに正確に個々人に対応しているかで判断される．これは，人を中心に考えれば，各個人がコミュニティを通していかに正確に認識されたかを評価することである．この状況を測る指標として認識率を導入する[15]．

いま，ある解法 $\mathcal{A}$ によって m 個のコミュニティ $(C_i, i = 1, .., m)$ が得られたとする．一方，実際には，x という名前を持つ人物が（D_0 の文書中に出現する範囲で）k 人存在していたとする（それぞれの人物を $p_j, j = 1, .., k$ とする）．このとき，β_{ij} を，コミュニティ C_i に属し，かつ，人物 p_j について言及している（当該人物を指し示している名前が記述されている）文書の数とすれば，

$$\beta_{i*} = \sum_{j=1}^{k} \beta_{ij}, \quad \beta_{*j} = \sum_{i=1}^{m} \beta_{ij}$$

はそれぞれコミュニティ C_i に属する文書の数と人物 p_j について言及している文書の数になる．人物 p_j に対して，

$$\frac{\beta_{ij}}{\beta_{i*}} > 0.5, \quad \frac{\beta_{ij}}{\beta_{*j}} > 0.5$$

という 2 つの条件が満たされるとき，p_j はコミュニティ C_i によって認識されたとし，これを $p_j \leftarrow C_i$ と表すことにする．なお，この条件の 0.5 という値は，それぞれの人物が 1 つのコミュニティによってのみ認識されるための最小値である．同姓同名の k 人のうち，首尾よくコミュニティによって認識された人物の割合を解法 $\mathcal{A}$ の**認識率**とする：

$$r(\mathcal{A}) = \frac{1}{k}\left|\{p_j|\ ^{\exists}i,\ p_j \leftarrow C_i\}\right|$$

厳密にいえば，この値は解法の特定の人名に対する有効性を示したもので，一般的な性能を示しているわけではない．そこで，解法の有効性は，複数の人名 $(x_i, i = 1, \ldots, h)$ に適用して得られるそれぞれの認識率の平均 $\overline{r}$ によって評価する．たとえば，一般語で構成したネットワークを連結性に着目して解析する解法は，

$$\overline{r} = \frac{1}{h}\sum_{i=1}^{h} r(\{G_{\mathcal{G}}(x_i, n_{max}),\ \mathcal{C}\})$$

の値で評価する．

4.3.2 評 価 結 果

文書集合 D_0 として，2003 年 7 月に主に jp ドメインのサイトを対象として Web ロボットにより収集した約 5 千万ページを用い，20 の人名 $(x_i, i = 1, 2, .., 20)$ について同姓同名分離を試みた．各 x_i と，この評価実験において識別されるべき同姓同名人物の数を**表 1** に示す．C_i と p_j の対応づけ，すなわち，各ページに出現する名前が実世界のどの人物に対応するかは，ページを実際に見て判定した．この判定作業は手間がかかるものなので，人名の選択においては非常に有名な人物を避け，作業者が有する知識で判定しやすいも

のであることを考慮した．その一方で，識別が困難と思われる場合 (x_{12}☆) や，ドラマや漫画の登場人物などの仮想的人物の名前 (x_2, x_{15}, x_{19}) をあえて含めた．個々の人名の出現傾向や認識率については，文献 15) を参照されたい．

図 4 は評価の結果をまとめたもので，特徴語数の増加にともなう認識率の平均 $\overline{r}$ の変化を，解法ごとに調べたものである．横軸には文書ごとの語数の平均 $\overline{n}$ をプロットした．用いた解法は，$\{\mathcal{G},\mathcal{C}\}$，$\{\mathcal{G},\mathcal{D}\}$，$\{\mathcal{P},\mathcal{C}\}$，$\{\mathcal{P},\mathcal{D}\}$ の 4 種類のすべてである．このグラフから，以下の 2 つの特徴が読み取れる．

まず，認識率は，解法 $\{\mathcal{P},\mathcal{D}\}$ では比較的複雑に変動しているが，それ以外では，基本的に単調に増加してピーク値に達し，その後，単調に減少している．$\{\mathcal{P},\mathcal{D}\}$ の場合も，おおまかにいえば，平均語数が 50 の近辺で最大値に達し，その前後でそれぞれ増加，減少傾向にある．また，いずれのネットワーク構成方法 ($\mathcal{G}$ または $\mathcal{P}$) においても，解析方法 ($\mathcal{C}$) を用いた場合は $\mathcal{D}$ の場合より，少ない平均語数でピークに達している．これは，連結性に着目した解析方法 ($\mathcal{C}$) はネットワークが 図 2 の (ii) の状態にあるとき有効であり，一方，稠密性に着目した方法はより辺の数が増えた状態，すなわち (iii) で，有効性が発揮されるであろうという 4.2 節での予想に合致した結果である．

次に，一般語を用いた場合 ($\mathcal{G}$) より人名を用いた場合 ($\mathcal{P}$) の方が，高い認識率が広い範囲で得られている．ピーク値で比較すると，$\{\mathcal{G},\mathcal{C}\}$ で 0.83 であったのに対し，$\{\mathcal{P},\mathcal{D}\}$ では 0.92 というより高い値が得られた．

表 1 評価に用いた人名
Table 1 Personal names used in the experiment.

i	x_i	同姓同名人物の数
1	伊庭幸人	1
2	上田次郎	3
3	江川卓	3
4	木下和彦	7
5	栗原はるみ	2
6	五斗進	1
7	五嶋みどり	1
8	新垣紀子	2
9	竹内郁雄	3
10	田中克己	16
11	中村紘子	4
12	野間佐和子	1
13	野村紀子	5
14	畑村洋太郎	1
15	菱沼聖子	2
16	福原愛	1
17	三浦麻子	5
18	水野晴郎	1
19	山岡士郎	2
20	和田英一	6

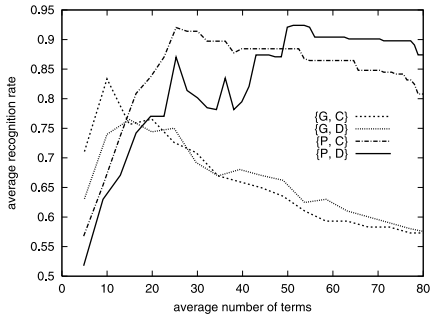

図 4 解法の妥当性の比較
Fig. 4 Comparison between solutions on the average recognition rate.

☆ 野間氏は講談社の社長であると同時に多くの組織の重要な役職に就いており，独立性の高い複数の文脈でその名前が出現することが多い．

5. 「よい」ネットワークの要件

5.1 構造的特徴

図 4 では，それぞれの解法の有効性を認識率の平均で示したが，個々のケース，すなわち，特定の人名 x に対する解法の有効性が平均と一致する（類似する）とは限らない．実際，平均とかなり異なる変動を示すものもある．**図 5** は 2 つの人名（x_A，x_B とする☆☆）の同姓同名分離を解法 $\{\mathcal{P},\mathcal{D}\}$ を用いて行った結果である．x_A の場合，平均語数が 50 の近辺で高い認識率に達した後その値が維持されており，平均的な認識率の変化と似通った挙動を示しているが，x_B ではそれとは異なった変化を示している．また，x_A のような，平均的な挙動を示す人名の間でも，認識率のピークの位置はすべてが一致するわけではない．

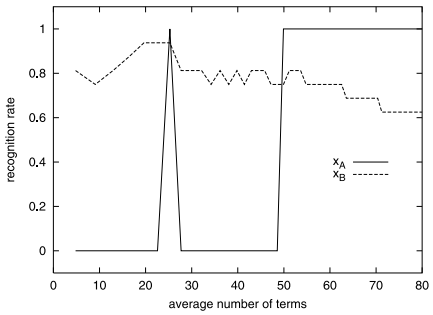

図 5 認識率における異なる変動
Fig. 5 Different styles of changes in the recognition rate.

☆☆ 具体的には，$A = 14$（畑村洋太郎），$B = 10$（田中克己）．

すなわち，一般論として，解法 $\{\mathcal{P}, \mathcal{D}\}$ が同姓同名人物を分離する方法として有効であることは分かったが，個々の人名 x について最善の結果を得るためには，さらに，平均語数（あるいは n_{max}）の調整がそれぞれ個別に必要になるのである．

そもそも，平均語数を調整するということは，ネットワーク構造を調整することにほかならない．つまり，$G_{\mathcal{P}}(x, n_{max})$ で n_{max} を変化させるということは，語数を介してネットワーク構造を調整しているということである．そこで，同姓同名分離という課題を解決する手段としてのネットワークの「よさ」を調べる方法として，語数というパラメータを用いるのではなく，直接構造上の特徴を調べるというアプローチが考えられる．最適な平均語数は，前述のように，人名によって異なるが,「よい」ネットワークの構造の特徴は異なる人名間で共有されている可能性がある．つまり，このアプローチは，人名に非依存な尺度でそれぞれの解法を評価できる可能性を持っている．

5.2 構造の比較

5.2.1 次 数 分 布

図 4 の評価結果が示すように，人名に基づくネットワーク $G_{\mathcal{P},\overline{n}}(x)$ を用いた方が，一般語を用いたネットワーク $G_{\mathcal{G},\overline{n}}(x)$ を用いた場合より高い認識率が得られるのであるから，この二者を比較し，前者に固有の特徴をさぐることが「よい」ネットワークの構造的特徴付けにつながると考えられる．

そこで，まず，2 種類のネットワーク $G_{\mathcal{G},\overline{n}}(x_A)$, $G_{\mathcal{P},\overline{n}}(x_A)$ の平均語数 $\overline{n}$ を $N = \{5, 10, 20, 30, 40, 50, 60, 70, 80\}$ の範囲で変化させ，次数分布（次数の累積確率分布）の変化を比較した．なお，人名 x_A は，図 5 の x_A と同一のものである．

図 6 の上方のグラフ (G) と下方のグラフ (P) がそれぞれ一般語を用いたネットワーク $G_{\mathcal{G},\overline{n}}(x_A)$ と人名を用いたネットワーク $G_{\mathcal{P},\overline{n}}(x_A)$ の平均語数 $\overline{n}$ の増加にともなう次数分布の変化を示したものである．同じ語数 $\overline{n}$ の場合を比較すると，一般語を用いたネットワークでは人名を用いた場合より多くの辺が生成されていることが分かる．さらに，一般語を用いたネットワークでは，語数の増加にともない，分布の形が変化している．語数が少ない範囲 ($\overline{n} \leq 10$) では次数の小さな（たとえば 5 以下の）頂点が存在しているが，その範囲を越えると，それらの頂点は存在しなくなる．一方，人名に基づくネットワークにおいては，$\overline{n} \in N$ の範囲では，つねに小さい次数の頂点が存在している．なお，これらの特徴は x_A 固有のものでなく，認識率の変動が異なる x_B においても，それぞれ同様な分布が得られている．

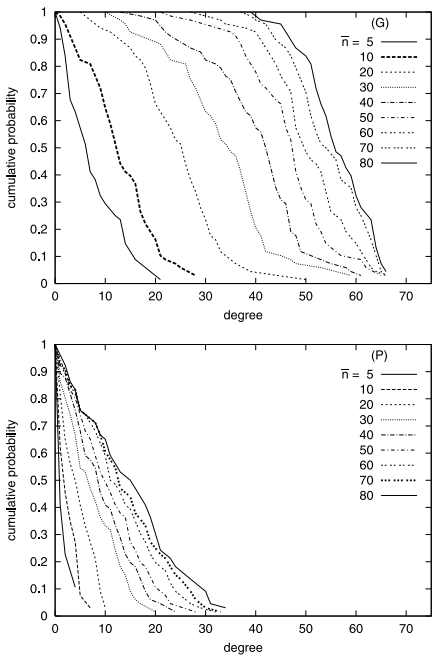

図 6 $G_{\mathcal{G},\overline{n}}(x_A)$ と $G_{\mathcal{P},\overline{n}}(x_A)$ の次数分布
Fig. 6 Degree distributions of $G_{\mathcal{G},\overline{n}}(x_A)$ and $G_{\mathcal{P},\overline{n}}(x_A)$.

このように，人名と一般語それぞれに基づくネットワークは，平均語数が同じでも，異なる次数分布を示すことが分かった．では，人名を用いたネットワークが示す次数分布，すなわち図 6 の (P) こそが,「よい」ネットワークの要件といえるのだろうか．図 6 をよく見ると，一般語に基づくネットワークでも，$\overline{n}$ の値が小さいときには人名の場合と類似した分布が認められる．たとえば，$G_{\mathcal{G},10}(x_A)$ と $G_{\mathcal{P},70}(x_A)$（それぞれ，図中太い線で示したもの）の分布は類似しており，辺の数もそれぞれ 419，428 とほぼ等しい．一方，認識率は，解法 $\mathcal{D}$ を使用した場合，それぞれ 0.6 と 1.0 と違いが生じている．この違いは次数分布では識別できない構造の違いに起因すると考えられる．

5.2.2 次数とクラスタ係数の相関

ネットワークの構造を示す統計量としては，次数分布以外にも，平均頂点間距離[4]，結合相関[18]，次数とクラスタ係数の関係[19] など数多く提案されている．ここでは，ネットワークの（局所的）稠密性との関連性がある，次数とクラスタ係数の関係を調べる．

図 7 は，辺の数がほぼ等しく，次数分布も類似している $G_{\mathcal{G},10}(x_A)$ と $G_{\mathcal{P},70}(x_A)$ それぞれについて，各頂点の次数とクラスタ係数の値をプロットしたものである（図中，それぞれ (G) と (P)）．(G) に比べて (P) には，より強い線形相関が見てとれる．実際，相関係

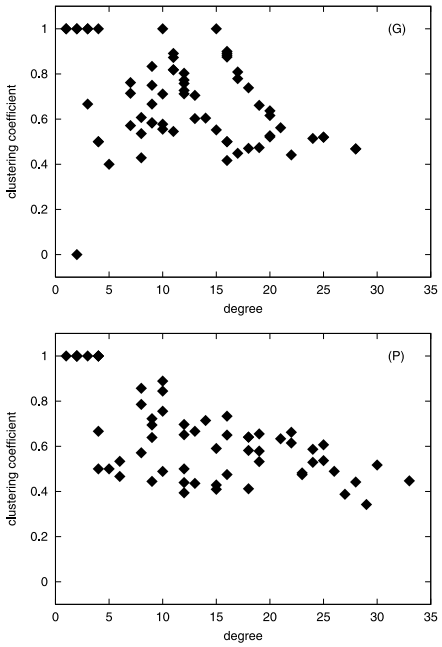

図 7 $G_{\mathcal{G},10}(x_A)$ と $G_{\mathcal{P},70}(x_A)$ における次数とクラスタ係数の関係
Fig. 7 Clustering coefficient vs. degree for $G_{\mathcal{G},10}(x_A)$ and $G_{\mathcal{P},70}(x_A)$.

数 c を計算すると，(G) では -0.31 であるのに対し，(P) では -0.63 であり，強い負の相関があることが分かる．

このことから，次数とクラスタ係数の相関の有無，すなわち c の値が「よい」ネットワークの指標になることが示唆される．この仮説を確かめるため，図 5 の例にもどり，2 種類のネットワーク（$G_{\mathcal{P},\overline{n}}(x_A)$ と $G_{\mathcal{P},\overline{n}}(x_B)$）の次数とクラスタ係数の関係を $\overline{n} = 30$ と $\overline{n} = 60$ という 2 つの場合について調べ，比較した．

図 5 を見ると，$G_{\mathcal{P},\overline{n}}(x_A)$ では $\overline{n} = 30$ のときよりも 60 の場合に高い認識率が得られている．一方，$G_{\mathcal{P},\overline{n}}(x_B)$ ではその逆で，$\overline{n} = 60$ の場合よりも 30 のときによい結果が得られている．よって，$G_{\mathcal{P},30}(x_A)$ に比べて $G_{\mathcal{P},60}(x_A)$ に，そして $G_{\mathcal{P},60}(x_B)$ よりも $G_{\mathcal{P},30}(x_B)$ に対して，次数とクラスタ係数の間により強い相関が期待される．**図 8** は，それぞれのネットワークの頂点ごとに次数とクラスタ係数をプロットしたものであるが，期待どおりの結果が得られている．それぞれの相関係数 c の値を**表 2** に示す．

さらに，$\overline{n} = 30$，60 の場合に限らず，人名 x_A，x_B それぞれについて認識率 r と相関係数 c の関連性を調べた結果を**図 9** に示す．ここでの c の値はみな負なので，グラフではその符合を逆にした $-c$ の値で相関の強さを示している．これらのグラフから，$-c$ の

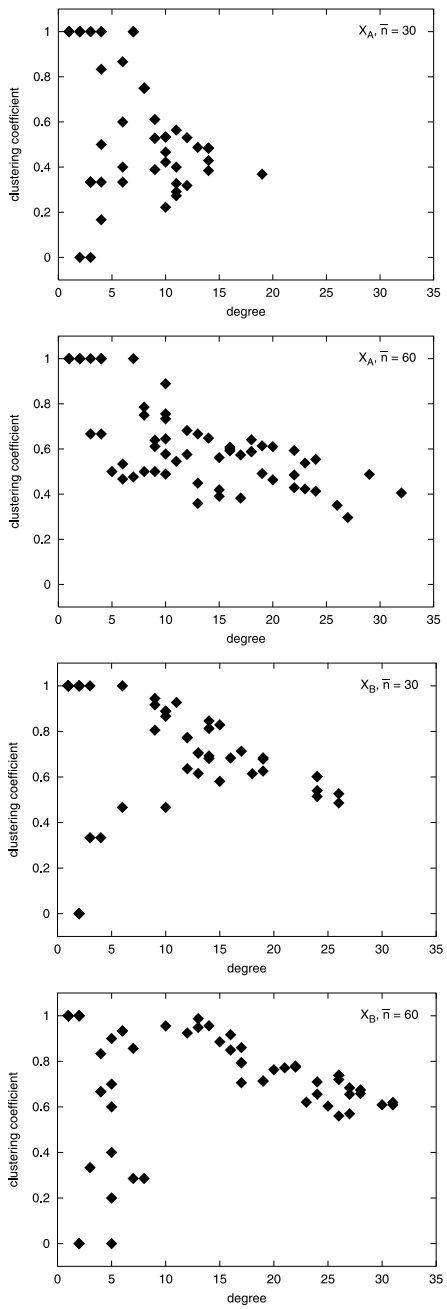

図 8 異なる $\overline{n}$ に対する次数とクラスタ係数の関係
Fig. 8 Clustering coefficient vs. degree for different values of $\overline{n}$.

値がおおむね認識率に連動していることが分かる．

この結果をふまえて，c の値に基づいた，「よい」ネットワーク，すなわち高い認識率を与えるネットワークの抽出を試みる．これは，たとえば，グラフの集合

$$\Omega = \{\, G_{\mathcal{P},\overline{n}}(x_i) \mid 1 \le i \le 20,\ \overline{n} \in N \,\}$$

に対して，認識率の平均

表 2 異なる $\overline{n}$ に対する次数とクラスタ係数の相関係数
Table 2 Correlation coefficient between the degree and the clustering coefficient for different values of $\overline{n}$.

	$\overline{n}$	c
x_A	30	−0.37
	60	−0.69
x_B	30	−0.06
	60	−0.31

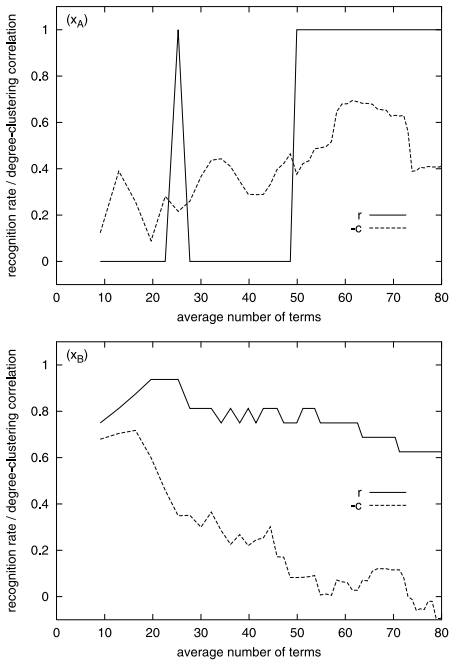

図 9 認識率 r と次数–クラスタ係数相関 c の関係
Fig. 9 Relationships between the recognition rate r and the correlation coefficient c.

$$r_* = \frac{1}{|\Omega|}\sum_{G\in\Omega} r(G)$$

が計算できるが，Ω の要素を c の値に基づいて取捨選択し，認識率の平均を向上させるという試みである．その方法を具体的に書くと，

$$\Omega_\lambda = \{\, G \in \Omega \mid c(G) < \lambda\},$$
$$r_\lambda = \frac{1}{|\Omega_\lambda|}\sum_{G\in\Omega_\lambda} r(G)$$

として，λ と r_λ の関係を調べる．もし，c が「よい」ネットワークを選ぶ指標となりうるのであれば，λ の値を小さくすることで r_λ を向上させることができるはずである．その結果は**表 3** のとおりであり，c は「よい」ネットワークの指標として妥当であることを示している．

表 3 c の値によるネットワークのふるい分け
Table 3 Picking up good networks based on c.

λ	$*$	−0.1	−0.4	−0.7	−0.8
r_λ	0.84	0.85	0.87	0.89	0.94

6. 考　　察

6.1 認識率の低下を招く要因

「よい」ネットワーク，すなわち高い認識率を与えるネットワークに共通した特徴として次数とクラスタ係数の間に線形相関がみられる（図 7 の (P)，図 8 の $x_A, \overline{n} = 60$ の場合など）ことから，その相関係数 c を「よい」ネットワークの指標として採用し，その妥当性を前章で確認した．一方，図 7 と図 8 のなかでこの相関が低いものに注目すると，全体的に相関が失われているのではなく，低次数の頂点におけるクラスタ係数の低下が原因で相関関係を阻害していることが読み取れる．この傾向は，とりわけ $G_{\mathcal{P},60}(x_B)$ で顕著である（図 8 最下部のグラフ）．このグラフでは，次数が高い範囲ではっきりとした線形相関が認められるにもかかわらず，低次数の範囲における低いクラスタ係数の影響で全体的な相関が失われている．

この，低次数におけるクラスタ係数の低下が認識率の低下に本質的に関わっているのであれば，全頂点の次数とクラスタ係数の相関をみる代わりに，次数の低い部分を調べることで低い認識率のネットワークを検出し，その排除により「よい」ネットワークを抽出する，というアプローチが考えられる．ただし，そのためには,「次数が低い部分」の恣意的でない定義を与えるためにも，さらに詳しい解析が必要である．

6.2 構造的特徴への翻訳による問題の単純化

本論文では,「よい」ネットワークを，それがネットワーク指向アプローチにおいて妥当（正確）な解析結果をもたらすものと定義し，その構造的特徴を探った．このアプローチの利点は，ネットワークに対する抽象的で複雑な要求条件（「妥当な解析結果をもたらす」ということ）をネットワークの構造的特徴に翻訳することで具体化し，問題を単純化できることにある．実際，5.2.2 項で示したように，語の種類（$\mathcal{P}$ と $\mathcal{G}$）の違いも，語の数（$\overline{n}$）の違いも，次数とクラスタ係数の相関という構造上の特徴に着目することにより統一的に比較できる．なお，本論文では要求条件の構造的特徴への翻訳可能性を暗に仮定し議論をすすめたが，一般には，このような翻訳が可能であるとは限らないことに注意が必要である．

7. ま と め

システムのつくりや振舞いを調べるための有力な手段として，要素間の相互関係をネットワークで表現し，その構造を解析するというアプローチがあるが，この方法により正確で詳細な解析結果を得ようとするならば，ネットワーク解析の方法だけでなく，ネットワークを構成する方法についても工夫する必要がある．いわば,「よい」ネットワークを構成する方法論が望まれるのだが，そのためには，まず，ネットワークの「よさ」を示す指標が必要である．

この観点から，本論文では，文書のネットワークを題材として,「よい」ネットワークの構造的特徴を探った．その結果，ネットワークの次数とクラスタ係数の間により強い負の相関が認められる場合に，より正確な解析結果が得られることが分かった．

従来,「よい」ネットワークを構成するためには，要素間の関係の定義や，関係の量を経験的に調整するという方法がとられてきた．これに対し，本論文で示した結果は，ネットワークの「よさ」をネットワーク自体の性質に基づいて議論できる可能性を示している．

参 考 文 献

1) Wasserman, S. and Faust, K.: *Social Network Analysis: Methods and Applications*, Cambridge University Press (1994).
2) Brin, S. and Page, L.: The anatomy of a large scale hypertextual Web search engine, *Proc. 7th International World Wide Web Conference*, pp.107–117 (1998).
3) Dean, J. and Henzinger, M.R.: Finding Related Pages in the World Wide Web, *Proc. 8th International World Wide Web Conference* (1999).
4) Watts, D.J. and Strogatz, S.H.: Collective dynamics of small-world networks, *Nature*, No.393, pp.440–442 (1998).
5) Barabási, A.-L. nad Albert, R.: Emergence of scaling in random networks, *Science*, No.286, pp.509–512 (1999).
6) Anderberg, M.R.: *Cluster Analysis for Applications*, Academic Press (1973).
7) Dunning, T.E.: Accurate Methods for the Statistics of Surprise and Coincidence, *Computational Linguistics*, Vol.19, No.1, pp.61–74 (1993).
8) Ng, H.T., Goh, W.B. and Low, K.L.: Feature selection, perceptron learning, and a usability case study for text categorization, *Proc. 20th ACM International Conference on Research and Development in Information Retrieval* (*SIGIR-97*), pp.67–73 (1997).
9) 相澤彰子，影浦　峡：著者キーワード中での共起に基づく専門用語間の関連度計算法，信学会論文誌，Vol.J83-D-I, No.11, pp.1154–1162 (2000).
10) 丹羽芳樹：動的な共起解析を用いた対話的文書検索支援，情処研究会報告，96-NL-115, pp.99–106 (1996).
11) Takano, A., Niwa, Y., Nishioka, S., Iwayama, M., Hisamitsu, T., Imaichi, O. and Sakurai, H.: Associative Information Access Using DualNAVI, *Proc. Kyoto International Conference on Digital Libraries 2000* (*ICDL'00*), pp.285–289 (2000).
12) Ohsawa, Y., Benson, N.E. and Yachida, M.: KeyGraph: Automatic Indexing by Cooccurrence Graph based on Building Construction Metaphor, *Proc. Advances in Digital Libraries Conference* (1998).
13) Matsuo, Y., Ohsawa, Y. and Ishizuka, M.: A Document as a Small World, *Proc. JSAI 2001, LNAI 2253*, pp.444–448, Springer-Verlag (2001).
14) Salton, G., McGill, M.J.: *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).
15) 佐藤進也，風間一洋，福田健介，村上健一郎：実世界指向 Web マイニングによる同姓同名人物の分離，情処論文誌：データベース，Vol.46, No.SIG8 (TOD26), pp.26–36 (2005).
16) Sato, S., Fukuda, K., Kazama, K. and Murakami, K.: Preliminary Results on Describing and Interpreting Context by Network Structure, *Proc. 1st International Workshop on Agent Network Dynamics and Intelligence* (*ANDI'05*), pp.39–44 (2005).
17) Girvan, M. and Newman, M.E.J.: Community structure in social and biological networks, *Proc. National Academy of Sciences USA*, Vol.99, No.12, pp.8271–8276 (2002).
18) Pastor-Satorras, R., Vázquez, A. and Vespignani, A.: Dynamical and correlation properties of the Internet, *Physical Review Letters*, Vol.87, No.25, 258701 (2001).
19) Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabási, A.-L.: Hierarchical organization of modularity in metabolic networks, *Science*, Vol.297, pp.1551–1555 (2002).

(平成 17 年 5 月 24 日受付)
(平成 18 年 1 月 6 日採録)

佐藤 進也（正会員）

1988 年東北大学大学院理学研究科数学専攻修士課程修了．同年日本電信電話（株）入社．協調作業における情報活用支援の研究に従事．現在，NTT 未来ねっと研究所主任研究員．ACM，Internet Society，電子情報通信学会各会員．

福田 健介

1999 年慶応義塾大学大学院理工学研究科計算機科学専攻後期博士課程修了．同年日本電信電話（株）入社．2002 年ボストン大学訪問研究員．2006 年 1 月より，国立情報学研究所情報基盤研究系助教授．インターネットトラフィックのダイナミクス，ネットワーク構造の統計的解析等の研究に従事．博士（工学）．

風間 一洋（正会員）

1988 年京都大学大学院工学研究科精密工学専攻修士課程修了．同年日本電信電話（株）入社．現在，NTT 未来ねっと研究所主任研究員．博士（情報学）．分散協調処理，情報検索の研究に従事．ソフトウェア科学会，ACM 各会員．

村上健一郎（正会員）

1979 年九州大学工学部情報工学科卒業．1981 年同大学院修士課程修了．同年日本電信電話公社入社．以来，超大型計算機用 OS，記号処理計算機，インターネットパラダイム，超高速インターネットプロトコルの研究に従事．2005 年 4 月より，法政大学ビジネススクールイノベーション・マネジメント研究科教授．博士（情報科学）．電子情報通信学会，ACM，ソフトウェア科学会各会員．主な著書『はやわかり TCP/IP』（共立出版，共著），『インターネット縦横無尽』（共立出版，共著），『インターネット』（岩波書店）等．