情報処理学会論文誌 Vol. 49 No. 6 2212–2223 (June 2008)

リンクに基づく分類のための ネットワーク構造を用いた属性生成

唐門 準[†1] 松尾 豊[†2] 石塚 満[†1]

近年，ネットワーク構造を持つデータを用いて学習や予測を行うためのさまざまな研究が行われている．ソーシャルネットワークや遺伝子のネットワークなど，ネットワーク構造を持つデータは多く，ネットワークからのデータマイニングは一般にリンクマイニングと呼ばれる．その中でも，リンクが張られている近傍ノードの情報も利用しながらノードの分類を行うタスクは「リンクに基づく分類」(link-based classification) と呼ばれ，その精度を上げるためにネットワーク構造を用いたさまざまな指標が考案されている．一方，これまで社会ネットワーク分析や複雑ネットワークの分野ではネットワークを評価する指標として，中心性，構造空隙，クラスタ係数などがよく用いられた．本稿では，この 2 つの研究の流れに注目し，従来から用いられてきた指標の生成を可能とするオペレータを定義し，リンクに基づく分類に適用する．論文のネットワークとソーシャルネットワークという 2 種類のデータに適用し，従来から用いられてきた指標の重要性を明らかにするとともに，未知の指標の可能性についても議論する．

Generating Social Network Features for Link-based Classification

Jun Karamon,[†1] Yutaka Matsuo[†2] and Mitsuru Ishizuka[†1]

There have been numerous attempts at the aggregation of attributes for relational data mining. Recently, an increasing number of studies have been undertaken to process social network data, partly because of the fact that so much social network data has become available. Among the various tasks in link mining, a popular task is *link-based classification*, by which samples are classified using the relations or links that are present among them. On the other hand, we sometimes employ traditional analytical methods in the field of social network analysis using e.g., centrality measures, structural holes, and network clustering. Through this study, we seek to bridge the gap between the aggregated features from the network data and traditional indices used in social network analysis. The notable feature of our algorithm is the ability to invent several indices that are well studied in sociology. We first define general operators that are applicable to an adjacent network. Then the combinations of the operators generate new features, some of which correspond to traditional indices, and others which are considered to be new. We apply our method for classification to two different datasets, thereby demonstrating the effectiveness of our approach.

1. はじめに

ウェブにおけるハイパーリンクやソーシャルネットワークサービスの知り合い関係は，ネットワークとしてとらえることができる．また，バイオサイエンスの分野でも遺伝子の相互作用や細胞におけるたんぱく質の相互作用などは，ネットワークとして取り扱うことができる[1]．このようなデータは，ノードが属性情報と関係情報の 2 種類の情報を持ち，ネットワーク構造を持つデータとして見なせる．こういったデータの関係情報に着目したマイニングは，最近では**リンクマイニング**と呼ばれる[⋆1]．リンクマイニングとは，リンク解析やウェブマイニング，関係学習，帰納論理プログラミング（ILP），グラフマイニングなどの複合領域として定義され，主なタスクとしては，リンク関係に基づくノードのクラスタリング，リンクに基づく分類，ノードのランキング，ノード解決（entity resolution），リンクの予測，サブグラフ発見などがある[2]．リンクに基づく分類（link-based classification）とは，リンクが張られている近傍ノードの情報も利用しながらノードの分類を行うタスクであり，確率伝播法や弛緩法，反復法などの代表的な手法が提案されている[3]．

一方，社会ネットワークに関する分析は古くから社会ネットワーク分析という社会学の一分野で行われており[4),5]，最近では，Web に関連してソーシャルネットワークサービスやブログ[6]，ソーシャルブックマーク[7] などを扱う研究もある[8]．ノードは actor（行為者），リンクは tie（紐帯）と呼ばれ，ネットワークやその中の個々のノード，あるいはリンクを特徴付けるための指標が考案されている[9),10]．たとえば，ネットワークの中で中心となる者は誰か（中心性の分析），個々のノードの役割は何か，また，誰と誰が競争関係にあり，誰

†1 東京大学大学院情報理工学系研究科
Graduate School of Information and Technology, The University of Tokyo
†2 東京大学大学院工学系研究科
Garduate School of Engineering, The University of Tokyo

⋆1 LinkKDD と呼ばれるワークショップが 2003 年から開催されており，また ACM SIGKDD の会誌である Explorations でも Link Mining の特集が組まれている[2]．

が効率的にネットワークを張っているか（構造同値，構造的空隙），ネットワーク上ではどのようなグループが構成されているか（クラスタ分析，クリーク分析）などの指標がある．これらの指標は 50 年以上にわたる社会学の分析に基づくものであり，実世界のネットワークを分析するのに有意義な指標である[11]．社会ネットワーク分析に比べて新しい複雑ネットワーク[12),13)] の研究でも，クラスタ係数（C）や平均パス長（L）などの指標がよく用いられる．

これまで，データマイニングの分野ではネットワークを分析するための多くの取り組みが行われてきた．たとえば，Backstrom らの研究では，コミュニティの発展に必要な要素が何かを分析している[14]．ユーザ（ノード）が所属するコミュニティを予測する問題に対して，コミュニティの情報を用いた 8 つの属性と，ノードの情報を用いた 6 つの属性を生成し，リンクに基づくノードの分類を行う．その結果，ユーザがあるコミュニティに所属する確率は，そのコミュニティ内に友人が多いほど高くなる傾向が見られた．さらに，コミュニティ内にいる友人が互いに知り合いである（つまりユーザ自身とトライアド[⋆1]を形成しているとき），ユーザはそのコミュニティに所属しやすい傾向が見られた．前者は自明だが後者は新たな発見であった．このように，リンクマイニングのタスクを扱ううえで，ネットワーク構造を用いた新たな属性を得ることは重要である．しかし，ネットワーク構造を用いた属性は Backstrom らの研究であげられているものがすべてではない．有用な属性を発見するには，ネットワーク構造を用いた属性を網羅的に生成する必要がある．

そこで本稿では，リンクマイニングのために有用な属性を体系的に生成するための手法を提案する．そのため，属性の生成過程を 3 つのステップに分割し，各ステップでいくつかの基本的なオペレータを定義する．各段階で定義されたオペレータを組み合わせることで，異なる属性を自動的に生成することが可能になる．生成された属性の一部は社会ネットワーク分析において用いられている属性と一致する．そのほかの属性は，これまでに用いられていない新たな属性である．生成された属性を用いて，リンクに基づく分類を行い，評価を行う．評価には，論文データベースである Cora のデータセットと，化粧品に関する女性向けのコミュニティサイトであるアットコスメのデータセットを用いた．

本稿の構成は以下のとおりである．まず 2 章では，本研究と関連する研究についていくつかの例をあげながら説明する．3 章では，社会ネットワーク分析における指標の詳細について概説する．4 章では，本研究での提案手法である，オペレータを用いた属性生成手法について説明する．5 章では，本提案手法を実際のデータセットに対して適用した結果について述べ，6 章で議論を行った後，7 章でまとめを述べる．

⋆1 3 者関係.

2. 関 連 研 究

本研究ではリンクに基づく分類を扱う．リンクに基づく分類は，機械学習における古典的な分類タスクにネットワークを活用するものであり，リンクマイニングの中でも基本的なタスクの 1 つである．

リンクに基づく分類では，リンク関係に基づき近傍のノードの情報も利用しながらノードの分類を行う．一般に次のように定義される．ネットワーク $\mathbf{G} = (\mathbf{V}, \mathbf{L})$ は，ノード集合 $\mathbf{V}$，ノード $x \in \mathbf{V}$ と $y \in \mathbf{V}$ の間にあるリンク l_{xy} の集合 $\mathbf{L}$ から構成される．各ノードには属性 a がありこれを $x.a$ のように表す．属性 a のとりうる値は $\mathbf{C} = (c_1, c_2, \ldots, c_n)$ である．このとき，属性 $x.a$ の値が与えられたネットワーク $\mathbf{G}_{train} = (\mathbf{V}_{train}, \mathbf{L}_{train})$ が与えられたときに，これから $\mathbf{G}_{test} = (\mathbf{V}_{test}, \mathbf{L}_{test})$ における各ノード $x \in \mathbf{V}_{test}$ の属性値 $x.a$ を推定する．ただし，$\mathbf{G}_{train}$ と $\mathbf{G}_{test}$ はそれぞれノードやリンクを共有しない異なるグラフである．

リンクに基づく分類のアルゴリズムの研究として，確率伝播法，弛緩法，反復法などが提案されている[3)]．たとえば，確率伝播法とは，観測された情報からの確率伝播によって，各ノードのラベルを更新していく方法である．ただしネットワーク中にループがないことを前提としており，ループがあっても適用可能なものとして，複結合ネットワーク確率伝播法（loopy belief propagation）がある．

ネットワーク構造を用いた属性生成に関する研究で最も関連が深いものは，前章でも述べた Backstrom らの研究[14] である．彼らは，大規模なブログホスティングサービスである LiveJournal[⋆2] と論文データベース DBLP の 2 つのデータセットを用いて，メンバあるいは論文の著者をノード，その友人関係または共著関係をリンクとしたネットワークをそれぞれ構築し，コミュニティの情報を用いた 8 つの属性と，**表 1** にあげたノードの情報を用いた 6 つの属性を生成し，各ノードをカテゴリへ分類することで，コミュニティの成長に必要な属性を分析した．その結果，ノードがあるコミュニティ，あるいは学会に所属する確率は，あるノードの隣接ノードでそのグループに所属しているノード数が多い方が上がるだけ

⋆2 ブログサービスがサービスを中心として，気に入った友人のリストの作成，自由に作られたコミュニティへの加入など，コミュニティシステムを持つサービスが提供されている．

表 1　Backstrom らの研究[14] で生成される属性の例
Table 1　Features used in Backstrom, et al. [14].

メンバ u とその友人のうちコミュニティ C に属するメンバの集合 S から生成される属性
コミュニティ C に属する友人の数（$\|S\|$）.
S 内のペアで直接のリンク関係を持つペアの数（$\|(u,v)\|u,v \in S \wedge (u,v) \in E_C\|$）.
S のうちリンク E_C により結ばれたペアの数.
リンク E_C で結ばれた友人間の平均距離.
リンク E_C でメンバ S から到達可能なコミュニティ C 内のメンバ数.
E_C で到達可能なメンバと S との平均距離.
E_C はコミュニティ C 内のリンク集合.

でなく，さらにそのような隣接ノードの間に直接のリンク関係がある方が上昇するという．このように，ノードの周辺のネットワーク構造を用いて，新たな属性を生成することで，リンクに基づくノードの分類に役立つ．

ところが，この研究では属性の生成はアドホックに行われており，たまたまいくつかの属性が有用であったものの，異なるドメインのネットワークでは用いられた属性が同じように有用であるとは限らない．そこで，ネットワーク構造を用いた有用な属性を幅広く得るためには，属性生成を体系化する必要がある．

このような研究として，Popescul らは，Statistical Relational Learning（SRL）において，リレーショナルデータベースにおける関係構造を得るために適切なクエリを用い，関係構造を用いた属性を生成する手法を提案している[15]．従来，SRL の研究領域では，属性間の関係を学習する Probabilistic Relational Models（PRM）の研究[16] がよく知られていた．PRM は，ベイジアンネットワークをより複雑な関係構造を扱える形に拡張したものであり，データベースが与えられたときに，リレーショナルスキーマとそれらに含まれるクラス内の各属性の確率的な依存関係を定義する．ここでは，個々のエンティティの属性だけを用いた分析ではなく，関係性を持つ（外部キーで参照されている）インスタンスの属性を用いた分析が行われている．

また Perlich らも Popescul らと同様に，関係データからの体系的な属性生成手法を提案している[17]．Perlich らの手法では，関係構造を複雑さの段階に応じたいくつかの階層に分類し，その階層に応じてリレーショナルスキーマや対象に依存する属性生成オペレータを導入している．NASDAQ における新規上場株の上場申請が受理されるかどうかを推定することで，提案手法の適用性と性能について論じている．

いずれも，本研究と目的は近いが，本研究では関係データよりもその総体としてのネットワークに着目しており，ネットワークデータへの適用可能性が高いこと，また社会ネットワーク分析における指標との関連を最大化していることなどが異なる点である．

3. 社会ネットワーク分析で用いられる指標

本章では社会ネットワーク分析で用いられる指標について概説する．これらの指標を体系的に生成するオペレータの定義を次章で述べる．

社会ネットワーク分析の分野では，古くからネットワークを評価するための多種の指標が提案されており，以下ではそれらの指標のうちよく知られた指標だけを取り上げる．ただし，ネットワーク内のノードの集合を N，ノード x における次数（リンク関係にあるノード数）を k_x，ノード x と y の距離を d_{xy} とする．

まず，社会ネットワーク分析における指標の中でも単純なものとして，ネットワーク密度がある．

ネットワーク密度　ネットワーク内に存在する各ノードのリンク具合を表すもので，$\frac{\sum_{x \in N} k_x}{N(N-1)}$ である．

次に，ネットワーク中での各ノードの影響の強さを測るものが中心性の指標であり，いくつかの算出方法がある[18]．

次数中心性　あるノードから他のノードに対して張られているリンクの数（を正規化したもの）である．$\frac{k_x}{N-1}$ で求められる．

近接中心性　ネットワーク中の特定のノードが他のノードにどれくらい容易に接近できる位置にいるかを表す指標で，$\frac{\sum_{x \neq y, y \in N} d_{xy}}{N-1}$ で表される．

媒介中心性　ネットワーク中の特定のノードが他のノードどうしの関係をどの程度媒介しているかを表す指標である．ノード y とノード z 間の最短パスの数を n_{yz}，そのうちノード x を通るノード y とノード z の最短パスの数を $n_{yz}(x)$ とすれば，$\sum_{y<z \in N} n_{yz}(x)/n_{yz}$ で求められる．

このほかの中心性の指標としてページランクとしても知られる**固有ベクトル中心性**がある．

また，近年複雑ネットワークの分野では，平均パス長，平均クラスタ係数などの指標が用いられる．以下ではこれら 2 つの指標について概説する．

平均パス長（L）　ネットワーク中のノード集合からすべてのノードペアの最短パス長の平

均である．

$$L = \frac{\sum_{x \in N, y \in N, x \neq y} d_{xy}}{N(N-1)}$$

で表される．

クラスタ係数（C）　ノード x に対して隣接するノード集合を E_x とすると，このノード集合 E_x の間で，どれくらいのリンクが張られているかを示すものである．これらの値をネットワーク中のすべてのノード N で平均した値を（平均）クラスタ係数 C と呼び，

$$C = \frac{1}{N} \sum_{x \in N} \frac{\sum_{y \in N_x, z \in N_x, y \neq z} a_{yz}}{N_x(N_x - 1)}$$

で求められる．ただし a_{xy} はノード x と y に直接のリンク関係があった場合に 1 を返しそれ以外の場合は 0 を返すもの，N_x はノード x に隣接するノード集合である．

さらに，構造同値や構造空隙もよく用いられる指標である．

構造同値　リンク関係に注目し，2 つのノードの役割の相違を表す指標であり，2 つのノードからのリンクが似たものであるほど値が小さくなる．2 つのノードそれぞれからリンク関係にある，2 つのノード集合のユークリッド距離をとることで求められる．たとえば，2 つのノードどうしがまったく同じノード集合にリンクを持っている場合，この値は 0 となり，ネットワーク上での 2 つのノードの役割はまったく同一であるといえる．

構造空隙　ネットワークにおける関係の分断のことを構造空隙という．ネットワーク上において 2 つの分断したクラスタが存在するとこれらのクラスタを結びつけるノードが存在すればそのノードは 2 つのクラスタを結びつけるという重要な役割を持つ．つまり互いに分断関係にあるノードを結びつけるノードほど構造空隙における評価値が高くなる．

社会ネットワーク分析や複雑ネットワーク分析などの分野で用いられるこれらの指標は，さまざまなネットワークデータにおける有用性が確かめられている．したがって，リンクに基づく分類を行ううえでも重要な属性になりうると考え，これを生成するオペレータについて次章で述べる．

4. 提 案 手 法

本章では，社会ネットワーク分析で用いられる指標をはじめ，ネットワーク構造を用いた属性を体系的に生成するための手法を提案する．

まず社会ネットワーク分析で用いられている指標を分析し，その生成をモデル化する．属性の生成をいくつかのステップに分解し，各ステップにおいて社会ネットワーク分析の指標生成に必要なオペレータを定義することで，これらの指標の生成をオペレータの組合せで実現することを考える．

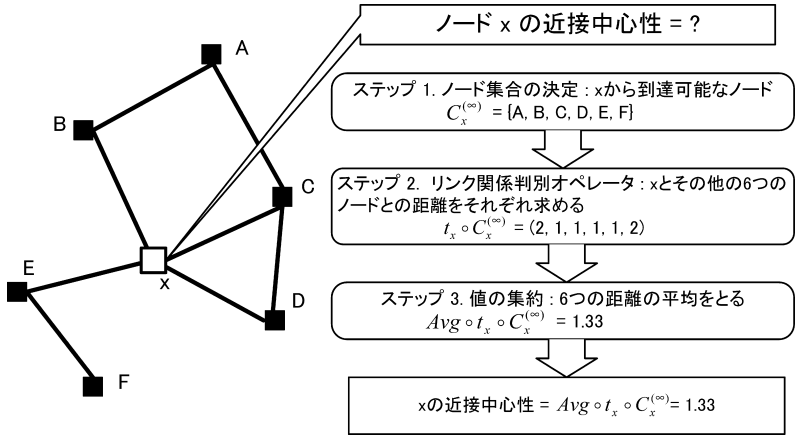

図 1　属性生成の流れ
Fig. 1　Flow of the feature generation.

では，ネットワーク構造を用いた属性を生成するにはどのようにオペレータを設計すればよいだろうか．ここでは**図 1** における近接中心性の計算を例にとって説明する．中心性の指標の生成過程は次のように分解することができる．まずステップ 1 として，中心性を求める対象ノード x から到達可能なノード集合を求める[⋆1]．ステップ 2 では，ノード x から第 1 段階で得られたノード集合の各ノードとの距離を求める．最後にステップ 3 として，得られた各距離を平均することで求める値が得られる．ステップ 1 からステップ 3 までの操作を行うオペレータをそれぞれ，$C_x^{(\infty)}$，t_x，Avg とおけば，ノード x における近接中心性は $Avg \circ t_x \circ C_x^{(\infty)}$ という 3 つのオペレータの組合せとして表現できる．

社会ネットワーク分析で定義された他の指標についても同様に分析を行った結果，本稿では属性の生成を次の 3 つのステップに分解し，各段階で必要なオペレータを定義することとする．

⋆1 近接中心性の計算は，ノード x を除いたすべてのノードを対象として行うものであるが，到達不可能ノードが 1 つでも存在すると近接中心性は無限大となってしまため，実質的には到達可能なノード集合を対象とするのは妥当である．

ステップ 1　対象ノードを決定する．

ステップ 2　ステップ 1 で得られたノード集合からノードペアの組合せをつくりノード間のリンク関係に関する何らかの値を調べる．

ステップ 3　ステップ 2 の結果を集計し属性値を得る．

3 つのステップのオペレータを組み合わせることで，さまざまな指標が得られる．以下では各ステップで定義されるオペレータについて説明する．

4.1　ステップ 1：ノード集合の決定

ステップ 1 ではノード集合を定義する．ノード集合を決めることで属性生成の対象とする部分グラフを得ることができる．

また本研究ではリンクに基づく分類を扱うため，ノードの属性値（分類したいカテゴリに関するカテゴリ属性）によるノード集合は重要である．たとえば，ノード x に隣接するノードのうち，あるカテゴリに属するノードに限定した場合のノード x の次数を得るには，ノードの属性によるノードの集合を定義する必要がある．そこで本稿では，ノード集合を求めるオペレータとして，以下に述べるように，距離に基づくオペレータとノードの属性値に基づくオペレータの 2 種類を定義する．

4.1.1　距離に基づくノード集合

距離に基づくノード集合とは，ノード x からの距離に基づいて決まるノード集合のことである．一例として x の隣接ノードは，ノード x から距離 1 のノード集合と同義である．同様にしてノード x から距離 2，距離 3 先のノード集合を得ることができる．このようなノード集合を得るオペレータを次のように定義する．

- $N_x^{(k)}$：ノード x から距離 k 離れたノード集合

ただし $N_x^{(0)}$ はノード x 自身を表す．

これを用いて一般にノード x から距離 k 以内にあるノード集合を得るオペレータを次のように定義する．

$$C_x^{(k)} = N_x^{(1)} \cup N_x^{(2)} \cup \ldots \cup N_x^{(k)}$$

4.1.2　属性値に基づくノード集合

属性値に基づくノード集合とは，ノードの持つ属性値が特定の値をとるノード集合のことである．たとえば，論文ネットワークにおいて，論文がある特定のカテゴリに所属する論文集合である．63 各ノードにはさまざまな属性が存在するが，本稿では特に分類の対象となるカテゴリ属性を考え，ノード集合の定義に用いる．カテゴリ属性が目的とする属性値をとるノード集合を「正のノード集合」と呼び，N_p と表す．また，正のノード集合 N_p と距離に基づくノード集合の積を考えることで，$C_x^{(k)} \cap N_p$ のようなノード集合を考えることができる．

このほかにも，ネットワーク中のすべてのノード集合 N や最大連結成分といった集合を定義することができる．しかし，これらの集合はどのノードにも同じ値をとるため，ここでは用いない．

4.2　ステップ 2：リンク関係判別オペレータ

本節ではステップ 1 で得られたノード集合に適用するオペレータを定義する．まず，2 つのノード間の関係を調べるオペレータを定義する．次にそれらを 3 つ以上のノード集合に対して適用できるように拡張する．本稿で定義するオペレータは次の 4 つである．

- $s^{(k)}(x, y)$：ノード x，y の間に距離 k 以内のリンク関係があるか
- $t(x, y)$：ノード x，y 間の距離
- $t_x(y)$：ノード x，y 間の距離（x との距離に限定）
- $u_x(y, z)$：ノード y，z の最短経路が x を経由するか

まず，$s^{(k)}(x, y)$ は，次のように定義される．

$$s^{(k)}(x,y) = \begin{cases} 1 & \text{if } x \text{ and } y \text{ are connected} \\ & \text{within } k \\ 0 & \text{otherwise} \end{cases}$$

たとえば $k = 1$ であれば，2 つのノード間に直接のリンク関係があるかどうかを調べるオペレータになる．また，$u_x(y, z)$ は，次のように定義される．

$$u_x(y,z) = \begin{cases} 1 & \text{if shortest path between } y \\ & \text{and } z \text{ includes } x \\ 0 & \text{otherwise} \end{cases}$$

ここまでは，2 つのノードに対して適用可能なオペレータを定義したが，これを 3 つ以上のノードを持つノード集合 N' に適用することを考える．具体的には次式のようにノード集合 N' から任意の 2 つのノードペアをつくり，それらに対して先に述べたオペレータを適用する．

$$\{operator(x, y) | x \in N', y \in N', x \neq y\}$$

例として，ノード集合 $\{n_1, n_2, n_3\}$ があったとき，このノード集合に関して直接のリン

表 2　オペレータリスト
Table 2　Operator list.

ステージ	表記	入力	出力	説明	適用レベル
1	$C_x^{(1)}$	node x	a nodeset	x の近接ノード集合	1
1	$C_x^{(\infty)}$	node x	a nodeset	x から到達可能なノード集合	2
1	$N_p \cap C_x^{(1)}$	node x	a nodeset	x の近接ノードのうち正のノード集合	3
1	$N_p \cap C_x^{(\infty)}$	node x	a nodeset	x から到達可能なノードのうち正のノード集合	3
2	$s^{(1)}$	a nodeset	a list of values	リンクがあれば 1，それ以外は 0	1
2	t	a nodeset	a list of values	ノードペア間のパス長	1
2	t_x	a nodeset	a list of values	ノード x とそのほかのノードの距離	2
2	u_x	a nodeset	a list of values	最短パスが x を経由していれば 1，それ以外は 0	2
3	Avg	a list of values	a value	平均	1
3	Sum	a list of values	a value	和	1
3	Min	a list of values	a value	最大値	1
3	Max	a list of values	a value	最小値	1
4	$ratio_p$	two values	value	すべてのノード集合（$C_x^{(k)}$）での値に対する正のノード集合（$N_p \cap C_x^{(k)}$）での値の割合	4

ク関係を調べるオペレータ $s^{(1)}$ を適用すると，$s^{(1)}(n_1, n_2)$，$s^{(1)}(n_1, n_3)$ と $s^{(1)}(n_2, n_2)$ を計算することになり，最終的にこれらの結果，値のリスト $(1, 0, 1)$ が得られる．

このような一連の処理を $s^{(1)} \circ N'$ のように表す．こうして，各オペレータを 3 つ以上のノード集合に適用可能にすることで，ステップ 2 の 4 種類のオペレータが定義される．

4.3　ステップ 3：値の集約

ステップ 3 では，ステップ 2 で得たリストを 1 つの値に集約するオペレータを定める．ステップ 2 で得たリストに対して，和（Sum），平均（Avg），最大値（Max），最小値（Min）をとるオペレータを考える．たとえば，ステップ 2 で $(1, 0, 1)$ のリストを得たとすると，このリストに対して Sum のオペレータを適用することで，2 という値を得ることができる．このようなステップ 1 から 3 に至る一連の操作を $Sum \circ s^{(1)} \circ N$ のように表す．ステップ 3 ではさらに分散や中央値などのオペレータなどが考えられるが，本稿では前記の 4 つのオペレータに限定する．

4.4　2 つの値の統合とオペレータの制限

3 つのステップのオペレータを順次適用することで値が得られるが，そういった 2 つの値を統合することも可能である．特に，リンクに基づく分類では，正のノード集合と全体のノード集合を比較することが重要であり，そのために特殊なオペレータである $ratio$ を付加的に用意する．

たとえば，ノード x に対して $Sum \circ t_x \circ C_x^{(1)}$ というオペレータ（次数に相当する）の値が 5 であるとする．また，正のノード集合 N_p だけに着目したオペレータ $Sum \circ t_x \circ (C_x^{(1)} \cap N_p)$ の値が 3 であるとすると，この 5 と 3 の値の比較は，近傍中での正のノードの数が多いかどうかの手がかりとなり，重要である．この 2 つの値の割合をとると，$3/5 = 0.6$ となる．このオペレータを $ratio$ と表す．

次に，本研究におけるオペレータの制限を述べる．本提案手法では多くのオペレータが適用されうる．たとえば，ステップ 1 におけるノード集合 $C_x^{(k)}$ を考えると，$k = \{1, 2, \cdots, \infty\}$ に対してノード集合は生成可能である．この大量の集合を候補とすることは計算時に大きな負荷となり，また多くの場合にそれほど意味のある値にならない．そのため，本研究では $k = \{1, \infty\}$ の 2 つの場合に限ってオペレータを定義する．ステップ 2 でも同様に，$s^{(k)}$ を最もシンプルなオペレータ $s^{(1)}$ のみに制限する．これらの制限は，必要に応じて緩めることが可能であるが，予備実験の範囲ではこれらの制限が計算量を下げながらも，性能の低下を小さく抑えることに役立っていることが分かっている．

本稿で用いるオペレータをまとめたものが，**表 2** である．ステップ 1〜3 に対して，それぞれ 4 つのオペレータを定義している．各ステップで 1 つずつオペレータを選択することで，$4 \times 4 \times 4 = 64$ のオペレータの組合せができる．さらに割合を考えることで $C_x^{(1)}$

と $N_p \cap C_x^{(1)}$ のノード集合をもとに求めた属性値の割合，$C_x^{(\infty)}$ と $N_p \cap C_x^{(\infty)}$ のノード集合をもとに得た属性値の割合を考えることができる．これらにより，各ノードに対して $64 + 32 = 96$ の属性を生成することができる．

これらのオペレータを用いて，社会ネットワーク分析で用いられる指標を生成することができる．以下にその例を示す．

- ネットワーク密度：$Avg \circ s^{(1)} \circ N$
- ネットワークの直径：$Max \circ t \circ N$
- 平均パス長：$Avg \circ t \circ N$
- 次数：$Sum \circ t_x \circ N_x^{(1)}$
- クラスタ係数：$Avg \circ s^{(1)} \circ N_x^{(1)}$
- 近接中心性：$Avg \circ t_x \circ C_x^{(\infty)}$
- 媒介中心性：$Sum \circ u_x \circ C_x^{(\infty)}$
- 構造空隙：$Avg \circ t \circ N_x^{(1)}$

また，次のように Backstrom らの研究[14] に含まれる属性を生成することも可能である．

- コミュニティ内の友人の数：$Sum \circ s^{(1)} \circ (C_x^{(1)} \cap N_p)$

このような指標を生成することができるのは，オペレータの設計時に生成可能となるように考慮したためだが，オペレータを組み合わせることにより，新たな属性を生成することが可能になる．たとえば，

- $Sum \circ t \circ C_x^{\infty}$
- $Sum \circ t_x \circ C_x^{\infty}$
- $Max \circ s^{(1)} \circ C_x^{(\infty)}$

のような属性を得ることができる．これらの属性はまだ知られていないが有用な属性になりうる．当然ながら，これらの新たな属性の中には属性として有用性が少ないものもある．たとえば，$Max \circ s^{(1)} \circ C_x^{(\infty)}$ は，到達可能なノード集合の中にリンクがあれば 1 を返す属性であるが，この属性は到達可能なノード集合があればつねに 1 をとるものであり，分類を行ううえで有用ではない可能性が高い．本研究では，生成された属性を用いて実際にリンクに基づく分類を行い，各属性の有用性について論じる．

5. 実験結果

本章では，提案手法を用いて生成した属性をもとにリンクに基づく分類を行い，提案手法の評価を行う．2 つのデータセットに対して適用する．

5.1 データセットと実験方法

提案手法の評価として，次の 2 つを行う．

(1)　提案手法がリンクに基づく分類において有用であるか．すなわち，オペレータで生成された属性によって分類精度が向上するか．

(2)　提案手法により生成された属性のうち，どの属性が分類に効果的であるか．

(1) の評価は，Backstrom らの研究[14] での評価手法を参考にして行った．まずあらかじめカテゴリを決め，そのカテゴリに属するノードを正例，属さないノードを負例とする．表 2 で定義したオペレータを用いて，各ノードに対して 96 の属性を生成し，これらの属性をもとに c4.5 法[19] を用いて決定木を学習し，各ノードが対象とするカテゴリに属するか属さないかを推定し，その再現率，適合率，F 値を評価する．ただし，定義したオペレータの有用性を示すため，表 2 に示すように，適用レベル 1〜4 まで段階的にオペレータを増やすこととした．はじめにレベル 1 のオペレータだけを用い，次にレベル 2 までのオペレータ，レベル 3 までのオペレータ，最後にすべてのオペレータを適用することで，順次，多くの属性を生成する．

(2) の評価は，(1) の評価で生成した決定木を用いて行う．決定木では上位に現れるほど，分類に有用な指標である．そこで決定木の上位に現れる属性ほど有用性が高くなるよう，深さ r に現れる属性に $1/r$ の点数をつけ，それらをすべてのカテゴリに関して足し合わせた値を各属性の有用性として評価した．

実験に用いたデータセットは，Cora データベースとアットコスメの 2 つである．以下ではこれらのデータセットの特徴とこれらのデータセットにおける実験手法について説明する．

5.1.1 Cora データセット

このデータセットは，McCallum ら[20] によって作られた Cora の論文データベースより作成した．Cora のデータベースは，コンピュータサイエンスの分野に属する約 30 万件の論文データを収集している．各論文は 69 の研究分野（カテゴリ）に分類されており，論文間の引用関係が与えられている．そのうちの 10 万件の論文はタイトルや，著者，ジャーナル，発表年などの詳細情報が付与されている．このデータを用いて論文をノード，論文間の引用関係をリンクとする論文ネットワークを構築した．ただしリンクは，すべて方向なしとした．

学習データとテストデータの生成は次のように行った．まず，対象とする研究分野を決定し，その研究分野に所属する論文あるいはその分野に所属している論文を引用している，ま

表 3　対象とした研究分野
Table 3　The research areas we select.

研究分野
/Artificial Intelligence/Knowledge Representation/
/Artificial Intelligence/Planning/
/Artificial Intelligence/Data Mining/
/Information Retrieval/Retrieval/
/Information Retrieval/Filtering/
/Artificial Intelligence/NLP/
/Databases/Object Oriented/
/Operating Systems/Distributed/
/Networking/Internet/
/Artificial Intelligence/Agents/
/Artificial Intelligence/Speech/
/Artificial Intelligence/Machine Learning/
Neural Networks/

たは引用されている論文集合をデータセットとした．この選択方法では負例は対象としているカテゴリに属していないにもかかわらず，そのカテゴリに属するノードに対してリンクを持っており，負例をランダムに選択するのに比べてより厳しい条件となっている．また，対象とする研究分野は，69 の研究分野からランダムに 5 分の 1 の研究分野を選択した．選択した論文のデータ集合は**表 3** のとおりである．

5.1.2　アットコスメデータセット[⋆1]

アットコスメとは 100 万人以上のメンバを持つ，女性向けとしては最大のコミュニティサイトである．サイト内で各ユーザは化粧品の推奨をしたり，感想を書いたりすることなどができる．アットコスメの特徴としては，各ユーザが気に入ったメンバをお気に入りメンバとして登録できる．またユーザはさまざまなコミュニティに所属することができる．各ユーザをノードとし，お気に入り関係をリンク（方向なし）とした社会ネットワークを構築する．タスクは，各ユーザを特定のコミュニティに所属するかしないかを分類することである．

学習データとテストセットの生成は先の Cora の論文データセットと同様に，カテゴリとして特定のコミュニティを指定し，そのコミュニティに所属するメンバをお気に入りリストに登録している，あるいは登録されているメンバの集合とした．コミュニティの選択は，所属メンバ数が 1,000 人以上いるという条件で行い，**表 4** に示した 12 のコミュニティを選択した．

⋆1 このデータセットは，個人情報をいっさい含まない形で研究用途に限り，アイスタイル株式会社より正式に提供を受けた．

表 4　対象としたコミュニティ
Table 4　The communities we select.

コミュニティ名
自然・低刺激派
スキンケアの鬼
外資ブランド好き
国産ブランド好き
安くていいもの好き
セルフチョイス派
メイク大好き！
カウンセリング派
ボディケア命
ネイル通
フレグランス好き
（ネット）通販好き

表 5　Cora のデータセットにおける再現率，適合率，F 値の変化
Table 5　Recall, precision and F-value in Cora dataset as adding operators.

	Recall	Precision	F-value
レベル 1	0.427	0.620	0.503
レベル 2	0.560	0.582	0.576
レベル 3	0.724	0.696	0.709
レベル 4	0.767	0.743	0.754

5.2　提案手法の有効性の評価

2 つのデータセットに対し，リンクに基づく分類タスクに対する再現率，適合率，F 値を，10 分割交差検定により調べた．**表 5** は，Cora データセットの「/Artificial Intelligence/Machine Learning/Neural Networks」の研究分野を対象に属性を生成し，分類精度を求めた結果である．1,682 のノード（論文）があり，この研究分野に所属するノード（正例）は 781 件である．この結果よりオペレータを増やすにつれて，F 値が向上していることが分かる．

決定木の上位ノードを示したものが**図 2** である．生成された決定木の深さは 8，属性数は 18 であり，図は適用レベル 2 の深さ 3 までのノードを示している．最上位ノード $Sum \circ s^{(1)} \circ C_x^{(\infty)}$ は，ノード x から到達可能なノード集合におけるリンクの数であり，意

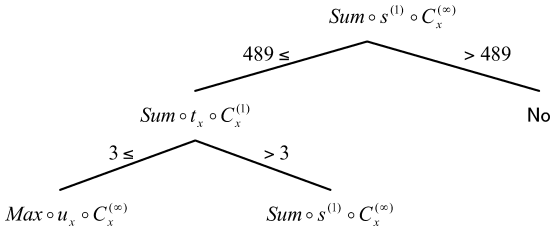

図 2　Cora データセットにおける適用レベル 2 で得られた深さ 3 までの決定木
Fig. 2　Top three levels of the decision tree using up to level 2 operators with the Cora dataset.

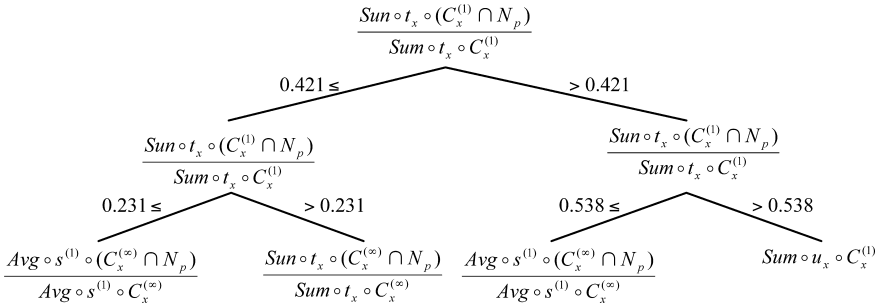

図 3　Cora データセットにおける適用レベル 4 で得られた深さ 3 までの決定木（オペレータ *ratio* は分母・分子の形で記述している）
Fig. 3　Top three levels of the decision tree using all operators with the Cora dataset.

味的にはノード x から到達可能なノード集合に限定したときのネットワーク密度に近い．深さ 2 に現れるノード $Sum \circ t_x \circ C_x^{(1)}$ は，ノード x の次数である．また，深さ 3 に現れる $Max \circ u_x \circ C_x^{(\infty)}$ は社会ネットワーク分析では使われない指標であり，ノード x を経由する最短パスが存在しているときに 1 をとり，そうでないときに 0 をとるような値である．

図 3 は，適用レベル 4 の決定木の深さ 3 までのノードである．このときの決定木の深さは 15，属性数は 48 であった．最上位のノード $\frac{Sum \circ t_x \circ (C_x^{(1)} \cap N_p)}{Sum \circ t_x \circ C_x^{(1)}}$ は，ノード x に隣接するノードの数に対する正のノードの数の割合である．深さ 3 には $\frac{Avg \circ s^{(1)} \circ (C_x^{(\infty)} \cap N_p)}{Avg \circ s^{(1)} \circ C_x^{(\infty)}}$ という属性があるが，これはノード x を含むサブグラフの密度である．また，$Sum \circ u_x \circ (C_x^{(1)} \cap N_p)$ は，ノード x の正の近接ノードにおける媒介中心性である．

表 6　アットコスメのデータセットにおける再現率，適合率，F 値の変化
Table 6　Recall, precision and F-value in the @cosme dataset as adding operators.

	Recall	Precision	F-value
レベル 1	0.419	0.555	0.473
レベル 2	0.544	0.629	0.580
レベル 3	0.707	0.745	0.722
レベル 4	0.731	0.757	0.742

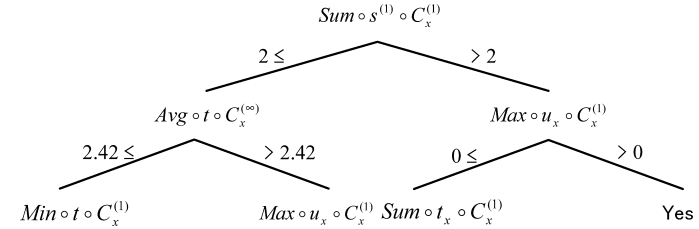

図 4　アットコスメのデータセットにおける適用レベル 2 で得られた深さ 3 までの決定木
Fig. 4　Top three levels of the decision tree using up to level 2 operators with the @cosme dataset.

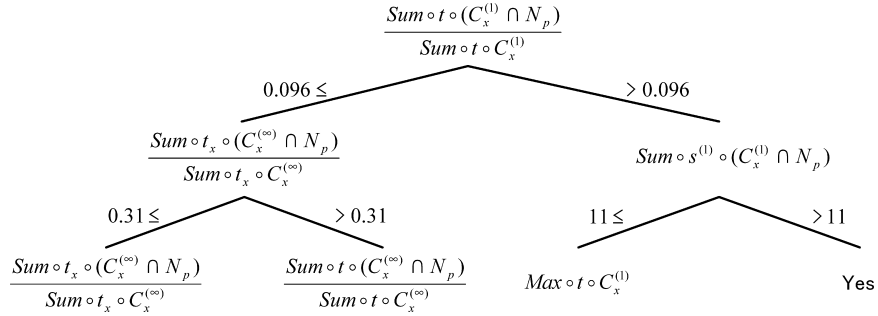

図 5　アットコスメのデータセットにおける適用レベル 4 で得られた深さ 3 まで決定木
Fig. 5　Top three levels of the decision tree using all operators with the @cosme dataset.

表 6 はアットコスメのデータセットにおける「スキンケアの鬼」のコミュニティに対して実験を行った結果である．ただし，データには 5,730 のノード（メンバ）があり，そのうちこのコミュニティに所属するノード（正例）は 2,807 件である．結果の傾向は Cora のデータセットと同様，オペレータを増やすに従い，再現率，適合率，F 値が向上している．また**図 4**，**図 5** はそれぞれ適用レベル 2，4 の際の決定木における上位ノードである．適用レベル 2 の決定木の深さは 13，属性数は 21 であり，適用レベル 4 の決定木の深さは 22，属

性数は 64 であった．図 4 の最上位ノード $Sum \circ s^{(1)} \circ C_x^{(1)}$ は，ノード x の近接ノード集合におけるリンクの数であり，クラスタ係数（$Ave \circ s^{(1)} \circ N_x^{(1)}$）に意味的に近い．また深さ 2 に現れるノード $Avg \circ t \circ C_x^{(\infty)}$ はノード x から到達可能なノード集合における平均パス長である．$Max \circ u_x \circ C_x^{(1)}$ は，ノード x が近接ノードペアのいずれかの最短パス上に存在していれば 1，そうでないならば 0 をとるような値，つまりクラスタ係数が 1 のとき 0，そうでないとき 1 になる．図 5 における最上位ノード $\frac{Sum \circ t \circ (C_x^{(1)} \cap N_p)}{Sum \circ t \circ C_x^{(1)}}$ は，隣接ノード集合における平均パス長に対する，正の隣接ノード集合における平均パス長の割合である．また深さ 2 のノード $\frac{Sum \circ t_x \circ (C_x^{(\infty)} \cap N_p)}{Sum \circ t_x \circ C_x^{(\infty)}}$ はノード x とノード x から到達可能なノードとの距離の和に対するノード x と正の到達可能なノードとの距離の和の割合である．同じく深さ 2 のノード $Sum \circ s^{(1)} \circ (C_x^{(1)} \cap N_p)$ は，隣接ノード集合におけるリンクの数であり，Backstrom らの研究で用いられた「トライアド関係の数」に相当する．深さ 3 のノード $Max \circ t \circ C_x^{(1)}$ は社会ネットワーク分析では用いられていない指標である．この属性はノード x と近接しているノード集合のあいだの距離の最大値であり，もしすべてのノードが直接つながっていれば 1，1 つでも直接のリンク関係がなければ 2 をとるような指標であり，ノード x のクラスタ係数に該当する指標となる．概して，正のノードの割合 $ratio$ を用いる属性に有用なものが多いことが分かる．

5.3　各属性の評価

Cora データセットとアットコスメのデータセットのそれぞれについて，平均して有用な属性を**表 7**，**表 8** に示す．決定木の深さ r に現れる属性の得点を $1/r$ として，全ケースの平均をとる．

さまざまな属性が分類に際して有効であるが，その中のいくつかはネットワーク密度（$Avg \circ s^{(1)} \circ C_x^{(\infty)}$）や，ノードの次数（$Sum \circ t_x \circ C_x^{(1)}$），媒介中心性（$Sum \circ u_x \circ (C_x^{(\infty)} \cap N_p)$）など，社会ネットワーク分析でよく知られた指標となっている．また Backstrom らの研究[14]で用いられている指標（$Sum \circ s^{(1)} \circ (C_x^{(1)} \cap N_p)$）も重要な指標であることが分かる．そのほかにも $Sum \circ s^{(1)} \circ C^{(1)}$ などいくつかの指標は社会ネットワーク分析ではあまり知られていない新しい指標であるが，これらの指標が示す意味は社会ネットワーク分析で古くから用いられている指標に近い（この例ではクラスタ係数（$Avg \circ s^{(1)} \circ N_x^{(1)}$）が近い）．

これらの結果から分かるように，社会ネットワーク分析で用いられている指標は一般的に有用であり，またそれ以外にも社会ネットワーク分析では通常用いられることの少ない有用な属性があるといえる．本提案手法は，体系的に属性を生成することができるので，ドメインにあわせて有用な属性を発見する目的に用いることができるのも特長である．

表 7　Cora のデータセットにおける上位 10 属性
Table 7　Top ten effective features in the Cora dataset.

順位	属性	説明	新規指標
1	$Sum \circ t_x \circ (C_x^{(1)} \cap N_p)$	ノード x の正の近接ノードの数.	
2	$Sum \circ t_x \circ C_x^{(1)}$	ノード x の近接ノードの数.	
3	$Avg \circ s^{(1)} \circ C_x^{(\infty)}$	ノード x から到達可能なノード集合内でのリンク数.	
4	$Sum \circ s^{(1)} \circ (C_x^{(1)} \cap N_p)$	ノード x に隣接する正のノード集合におけるリンク数[14].	
5	$Max \circ t \circ (C_x^{(1)} \cap N_p)$	ノード x の正の近接ノード集合における直径.	
6	$Sum \circ s^{(1)} \circ C_x^{(1)}$	ノード x に隣接するノード集合におけるリンク数.	○
7	$Sum \circ s^{(1)} \circ C_x^{(\infty)}$	ノード x から到達可能なノード集合内でのリンク数.	
8	$Max \circ u_x \circ (C_x^{(\infty)} \cap N_p)$	ノード x を経由する最短パスがあるか.	○
9	$Max \circ s^{(1)} \circ (C_x^{(1)} \cap N_p)$	x と 2 つの正の近接ノードの間にトライアド関係があるか.	○
10	$Avg \circ s^{(1)} \circ C_x^{(\infty)}$	ノード x から到達可能なノード集合内でのネットワーク密度.	

表 8　アットコスメのデータセットにおける上位 10 属性
Table 8　Top ten effective features in the @cosme dataset.

順位	属性	説明	新規指標
1	$Sum \circ s^{(1)} \circ (C_x^{(\infty)} \cap N_p)$	ノード x から到達可能なノード集合内でのリンク数.	
2	$Sum \circ s^{(1)} \circ C_x^{(1)}$	ノード x に隣接するノード集合におけるリンク数.	○
3	$Sum \circ t_x \circ C_x^{(1)}$	ノード x の近接ノードの数.	
4	$Avg \circ t \circ (C_x^{(\infty)} \cap N_p)$	ノード x から到達可能な正のノード集合における平均パス長.	
5	$Sum \circ u_x \circ (C_x^{(\infty)} \cap N_p)$	ノード x の媒介中心性.	
6	$Avg \circ t \circ C_x^{(\infty)}$	ノード x から到達可能なノード集合における平均パス長.	
7	$Avg \circ s^{(1)} \circ C_x^{(\infty)}$	ノード x から到達可能なノード集合内でのネットワーク密度.	
8	$Avg \circ s^{(1)} \circ (C_x^{(\infty)} \cap N_p)$	ノード x から到達可能な正のノード集合内でのネットワーク密度.	
9	$Sum \circ t_x \circ (C_x^{(\infty)} \cap N_p)$	ノード x から到達可能な正のノード集合における近接中心性.	
10	$Avg \circ u_x \circ C_x^{(1)}$	ノード x に隣接するノード集合における媒介中心性.	

6. 議　　論

本研究で定義したオペレータ以外にも，他のオペレータを付加的に用いることで，提案手法をさらに拡張することができる．たとえば，

- 中心化：e.g., $Max_{n \in N} \circ Sum \circ s^{(1)} \circ C_x^{(\infty)} - Avg_{n \in N} \circ Sum \circ s^{(1)} \circ C_x^{(\infty)}$
- 平均クラスタ係数：$Avg_{n \in N} \circ Avg \circ s^{(1)} \circ N$

などは，$Avg_{n \in N}$ というオペレータ（ノード全体に平均をとる）を定義することで実現で

きる．そのほか，たとえば，2つのノード間の距離をランダムサーファーをひきつける確率によって求めるオペレータとして定義することで，固有ベクトル中心性を求めることも可能である．ただしオペレータの実装が複雑になり計算コストが高い点が問題となる．また，本研究ではリンクの方向を扱っていないが，これは社会ネットワークの指標の多くがリンクの方向を考慮していないためである．リンクの方向を考慮した拡張も今後の拡張として必要であろう．

本稿で定義したオペレータはあくまでネットワーク構造を用いた属性を体系的に生成するための手法の可能性を示すために定義したものであり，必ずしもこれらのオペレータが最適かつ有用であると結論づけることはできない．どのようなオペレータの組合せがどういったタスクに効果的であるかは，今後，分析を進めていきながら明らかにすべき課題である．

本稿では，特にリンクに基づく分類タスクに焦点をあてて実験を行った．しかし，原理的には，本提案手法はさまざまなリンクマイニングのタスクに適用可能である．たとえば，リンク予測への適用を考える．リンク予測とは，2つのノード x と y が与えられたときにそのノード間にリンクが発生するかを予測する問題であるので，次のようなオペレータが必要となる．

- 2つのノード x，y の属性値の集約を行うオペレータ
- 2つのノードに共通する近接ノード集合（$C_x^{(k)} \cap C_y^{(k)}$）を得るオペレータ

今後は他のタスクへの適用を行いながら，社会ネットワークのマイニングのための一般的な属性生成の方法を構築したいと考えている．

7. ま と め

本研究では，データマイニングと社会学の間のギャップを埋めるために必要な研究として，社会ネットワーク分析で用いられている指標を体系的に生成する手法を提案した．提案手法では属性生成の過程を3つのステップにわけ，各ステップでオペレータを定義し，それらのオペレータの組合せにより属性を生成した．またこの手法を Cora とアットコスメの2つのデータセットに適用し，リンクに基づくノードの分類への有効性を示した．2つのデータセットを用いた実験を通して，この2つのデータセットに対する結果の傾向が似ていること，また中心性やネットワーク密度など社会ネットワーク分析で用いられている指標が有用であることが分かった．*ratio* というオペレータは特に有用であり，社会学の分野では用いられていないが，タスクによっては有用な属性であることが明らかになった．

ネットワークと機械学習の分野は，徐々にその融合領域での研究が進んでおり，Web やバイオサイエンスにおいて必要性が高まっている．本研究が1つの重要な知見を提供することになれば，著者らの幸いとするところである．

謝辞　本研究の実験にあたり，データの研究用途での利用にご協力いただいたアイスタイル株式会社および成蹊大学の山本晶先生に心より感謝いたします．

参 考 文 献

1) Barabási, A.-L.：新ネットワーク思考，NHK 出版 (2002).
2) Getoor, L. and Diehl, C.P.: Link Mining: A survey, *SIGKDD Explorations*, Vol.2, No.7 (2005).
3) Sen, P. and Getoor, L.: Link-based Classification, Technical Report CS-TR-4858, University of Maryland (2007).
4) Wasserman, S. and Faust, K.: *Social network analysis–Methods and Applications*, Cambridge University Press, Cambridge (1994).
5) Scott, J.: *Social Network Analysis: A Handbook, 2nd ed.*, SAGE publications (2000).
6) Adamic, L. and Glance, N.: The Political Blogosphere and the 2004 U.S. Election: Divided They Blog, *LinkKDD-2005* (2005).
7) Golder, S. and Huberman, B.A.: The Structure of Collaborative Tagging Systems, *Journal of Information Science* (2006).
8) 松尾　豊：Web2.0 時代の個人とコラボレーション，情報処理，Vol.47, No.11 (2006).
9) 安田　雪：実践ネットワーク分析，新曜社 (2001).
10) 金光　淳：社会ネットワーク分析の基礎—社会的関係資本論にむけて，勁草書店 (2003).
11) 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).
12) Watts, D.: *Six Degrees: The Science of a Connected Age*, W.W. Norton & Company (2003).
13) Barabási, A.-L.: *LINKED: The New Science of Networks*, Perseus Publishing, Cambride, MA (2002).
14) Backstrom, L., Huttenlocher, D., Lan, X. and Kleinberg, J.: Group formation in large social networks: Membership–Growth, and Evolution, *Proc. SIGKDD'06* (2006).
15) Popescul, A. and Ungar, L.: Statistical Relational Learning for Link Prediction, *IJCAI-03 Workshop on Learning Statistical Models from Relational Data* (2003).
16) Friedman, N., Getoor, L., Koller, D. and Pfeffer, A.: Learning Probabilistic Relational Models, *Proc. IJCAI-99*, pp.1300–1309 (1999).
17) Perlich, C. and Provost, F.: Aggregation Based Feature Invention and Relational Concept Classes, *Proc. KDD 2003* (2003).
18) Freeman, L.C.: Centrality in social networks: Conceptual clarification, *Social Net-*

works, Vol.1, pp.215–239 (1979).
19) Quinlan, J.R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann, California (1993).
20) McCallum, A., Nigam, K., Rennie, J. and Seymore, K.: Automating the Construction of Internet Portals with Machine Learning, *Information Retrieval Journal*, Vol.3, pp.127–163 (2000). www.research.whizbang.com/data

(平成 20 年 1 月 2 日受付)
(平成 20 年 3 月 14 日採録)

唐門　準

2006 年東京大学工学部電子情報工学科卒業．2008 年同大学院情報理工学系研究科電子情報学専攻修士課程修了．人工知能，Web マイニング等に興味がある．

松尾　豊（正会員）

1997 年東京大学工学部電子情報工学科卒業．2002 年同大学院博士課程修了．博士（工学）．同年より，産業技術総合研究所研究員．2005 年 10 月よりスタンフォード大学客員研究員．2007 年 10 月より，東京大学大学院工学系研究科総合研究機構准教授．人工知能と Web マイニングに興味がある．人工知能学会，言語処理学会，AAAI の各会員．

石塚　満（正会員）

1971 年東京大学工学部電子工学科卒業，1976 年同大学院博士課程修了．工学博士．同年 NTT 入社，横須賀研究所勤務．1978 年東京大学生産技術研究所・助教授（1980〜1981 年 Purdue 大学客員准教授），1992 年東京大学工学部電子情報工学科教授，2001 年情報理工学系研究科電子情報学専攻，2005 年同創造情報学専攻（電子情報学専攻兼任）．研究分野は人工知能，Web インテリジェンス，次世代 Web 情報基盤，生命的エージェントによるマルチモーダルメディア．IEEE，AAAI，人工知能学会（前会長），電子情報通信学会，映像情報メディア学会，画像電子学会等の各会員．