Vol. 47 No. 4 情報処理学会論文誌 Apr. 2006

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[†] 松尾 豊[†] 石塚 満[†]

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ている．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ための指標として，中心性，構造空隙，クラスタ係数などがよく用いられる．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を行う．論文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し，さら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,[†] Yutaka Matsuo[†] and Mitsuru Ishizuka[†]

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 (SNS) の知り合い関係は，ネットワークとして捉えることができる．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[4]．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みなせる．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**と呼ばれることもある☆．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[7]．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，確率伝搬法や弛緩法，反復法などの代表的な手法が提案されている[14]．

一方，社会ネットワークに関する分析は古くから社

† 東京大学
The University of Tokyo

☆ 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[7]．

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[13),15)]，最近では，Web に関連して SNS やブログ[1)]，ソーシャルブックマーク[8)] などを扱う研究もある[20)]．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，あるいはエッジを特徴付けるための指標が考案されている[17),19)]．例えば，ネットワークの中で中心となる者は誰か (中心性の分析)，個々のネットワーク上における役割は何か (役割の分析)，また，誰と誰が競争関係にあり，誰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) などの指標が挙げられる．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とされている[18)]．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[3),16)] の研究でも，クラスタ係数 (C) や平均パス長 (L)，リンクの次数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．例えば，Backstrom らの研究では，社会でグループや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[2)]．社会ネットワークにおいて所属するコミュニティを予測する問題を対象とし，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ほうが☆，ほうが，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．リンクに基づく分類をはじめ，リンクマイニングのタスクを扱う上で，ネットワーク構造を用いた新たな属性を作ることは重要であると考えられる．しかし，ネットワーク構造を用いた有益な属性は Backstrom らの研究で挙げられている属性以外にも存在しうる．有益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ことが必要になる．

そこで本稿では，社会ネットワーク分析で用いられている指標をはじめ，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，各ステップではいくつかの基本的なオペレータを定義する．そして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中心性などの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．その他の属性は，これまでに用いられていない新たな属性となる．さらに，生成された属性を用いて，リンクに基づく分類を行う．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に対して適用し，提案手法の有用性を示す．

本稿の構成は以下のようになっている．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挙げながら説明する．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説明する．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概説する．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際の実験結果について説明し，まとめと今後の結論について言及する．

2. 関 連 研 究

本研究ではリンクに基づく分類タスクを扱う．なぜならば，分類タスクはデータマイニングにおける単純なタスクのひとつであり，リンクを用いることによる効果を測るのに最適であると考えられるからである．リンクに基づく分類タスクとは，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タスクであり，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 $\mathbf{C} = (c_1, c_2, ..., c_n)$ 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 $x.a \in \mathbf{V}_{test}$ を推定するものである．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とする．

リンクに基づくノードの分類アルゴリズムの研究として，確率伝播法，弛緩法，反復法[14)] など数多く行われている．例えば，確率伝搬法とは，観測された情報からの確率伝搬によって，各ノードのラベルを更新していく方法である．ただしこの手法はネットワーク中にループしたパスがないことを前提としており，そのような条件下に適用可能な確率伝搬法として，複結合

☆ 自分と 2 人の友人との間に三者関係が成立しているとき，これを「トライアド関係」という．

ネットワーク確率伝播法 (loopy belief propagation) が提案されている．

ネットワーク構造を用いた属性生成に関する研究としては Backstrom らの研究[2] が挙げられる．彼らは，大規模なブログホスティングサービスである LiveJournal☆と論文データベース DBLP の 2 つのデータセットを用いて，メンバー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エッジ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発見している．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ほうが上がるだけでなく，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ほうが上昇するという．このように，ノードの周りに構築されたネットワーク構造を用いて，新たな属性を生成することは，リンクに基づくノードの分類に役立つと考えられる．

ところが，この研究では属性の生成は人手を介して行われており，異なるドメインのネットワークでは有益な属性を必ずしも得られるとは限らず，有益な属性が他にも存在する可能性もある．そこで，ネットワーク構造を用いた有益な属性を得るためにはその属性生成手法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考え，それらを用いて関係構造を用いた属性を生成する手法を提案している[11]．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[6] 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，関係性をもつ (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．このように，関係性は分析における重要な指標となると考えられる．しかし，この学習モデルは属性選択を考えたものであり，属性生成は人手を介して行われていた．そこで，関係データから体系的に関係構造を用いた属性を生成するための手法を提案したのがこの研究であり，提案手法を用いて論文の引用関係や著者，学会情報を持つ Citeseer のデータで，論文の参照リンクを推定することで，手法の有用性を示している．

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[10]．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．さらに提案手法を用いて 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，本提案手法の適用性と性能について論じている．

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．

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様々な指標の研究がなされており，以下ではそれらの指標のうちよく知られた指標について説明する．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，$\frac{\sum_{x \in N} k_x}{N(N-1)}$ として求められる．

また，ネットワーク分析においてよく用いられる指標として中心性の指標がある．例えば，SNS における人間関係を考えたとき，他者とのつながりが多い人ほどそのネットワークでの影響度が大きいと考えられる．このように，ネットワーク中での各ノードの力の強さが中心性であり，いくつかの算出方法がある[5]．以下ではそのうち本稿で用いる指標について概説する☆☆．

次数中心性 ノードの次数とはあるノードから，他のノードに対して張られているリンクの数である．つまり，次数中心性とは各ノードがどれくらい他のノードと関わりを持っているかを表す指標である．$\frac{k_x}{N-1}$ 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

☆ 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．はてなダイアリーに近いサービスであるとされる．

☆☆ この他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．

表 1 Backstrom らの研究[2] で生成される属性の例.

メンバー u とその友人のうちコミュニティC に属するメンバーの集合 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 ($\|\|(u,v)\|u,v \in S \wedge (u,v) \in E_C\|$).(ただし E_C はコミュニティC 内のエッジ)
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 でメンバー S から到達可能なコミュニティC 内のメンバー数.
E_C で到達可能なメンバーと S との平均距離.

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同士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．

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．この値が高いほど知り合いとの間でトライアド関係が構築されやすい．特に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，

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

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 を返すものとする．

この他にも，構造同値，構造空隙をはじめとする様々な指標が提案されている．

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 つのノードのリンクが似たものほど値が小さくなる．二つのノードのリンク関係のユークリッド距離をとることで求められる．例えば，2 つのノード同士がまったく同じノード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の物であるといえる．

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においてはこのほかにも様々な指標が提案されているが[13),15)]，本章では特に本研究で生成対象とする指標について説明した．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，古くからネットワークを分析する上での有益性が示されており，リンクに基づく分類を行う上でも重要な属性になりうると著者らは考える．

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．

社会ネットワーク分析で用いられる指標を，各ノードの属性として生成することは有用であると考えられる．また Backstrom らの研究[2)] が示すように，社会ネットワーク分析で用いられている指標以外にも新たにネットワーク構造を用いた重要な属性があることが示されている．ところがこの研究は，個々の属性を生成して，その属性の有益性を示したものであり，ネットワーク構造を用いた有益な属性はこのほかにも存在する可能性がある．有益な属性をできるだけ発見するには，ネットワーク属性を網羅的に生成し，一つ一つ検証することが求められる．そこでネットワーク構造を用いた属性生成を体系化し，網羅的に属性を生成する手法が必要となる．

そこで本稿では，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，その生成過程をモデル化する．生成過程をいくつかの過程に分解し，各過程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

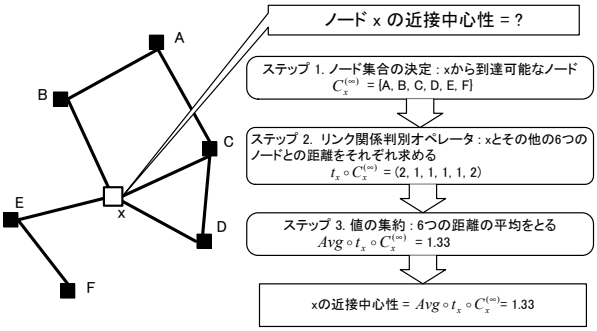


図 1 近接中心性の計算

レータの組み合わせで実現しようと考える．

ではネットワーク構造を用いた属性を効率よく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．まず第一段階として，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☆．第二段階では，ノード x から第一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．最後に第三段階として，得られた各距離を平均することで求める値が得られる．第一段階から第三段階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み合わ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過程を次の 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．

ステップ 1 対象ノードを決定するオペレータを定義する．

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み合わせをつくりノード間のリンク関係を調べるオペレータを定義する．

ステップ 3 ステップ 2 の結果を集計しネットワーク構造を用いた属性を得るオペレータを定義する．

3 ステップのオペレータを組み合わせることで，社会ネットワーク分析で用いられる指標が得られる．さらにオペレータの組み合わせによっては，新たな指標を作り出すことができる．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．

☆ 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，この理論では，到達不可能ノードがひとつでも存在すると近接中心性は無限大となってしまう．そこで本稿では到達可能なノード集合を対象としている．

4.1 ノード集合の決定

本節ではステップ 1 のノード集合を決定するオペレータを定義する．ノード集合を決めることにより属性生成の対象とするサブグラフを得ることができる．例えば，ノード x の次数は x の隣接ノード数なので，x に隣接するノード集合を考える必要がある．

また本研究ではリンクに基づく分類タスクを扱うため，ノードの属性値 (カテゴリ属性) によるノード集合は重要だと考えられる．例えばノード 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にノード x の次数がどうなるかを考えることが可能になる．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 種類を定義することとする．以下ではこの 2 つについて説明する．

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．

- $N^{(k)}(x)$: 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．

$$C^{(k)}(x) = N^{(1)}(x) \cap N^{(2)}(x) \cap \ldots \cap N^{(k)}(x) \quad (1)$$

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．例えば，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を考えることができる．ただし各ノードには様々な属性が存在するため，本稿では特にカテゴリ属性を重要と考えノード集合の決定に用いる．こうして得られたノード集合を「正のノード集合」と呼び，N_p と表す．

このような正のノード集合 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☆☆．以下ではこのようなノード集合 $C_x^{(k)} \cap N_p$ を,「属性値に基づくノード集合」とする．

このほかにもネットワーク中のすべてのノード集合

☆☆ この集合 N_p と，距離に基づくノード集合 $C_x^{(k)}$ との間で，AND/OR/NOT の真偽を考えることで，16 通りのノード集合が考えられる．

N といった集合を定義することができる．ただし N を元に生成される属性は各ノードとも同じ値をとるため，分類問題ではこの属性を用いない．

4.2 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 つのノード間にある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．ただし本稿で定義するオペレータを次の 4 つに限定する．

- $s^{(k)}(x,y)$：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

以下ではこれらのオペレータの詳細について説明する．

まず，$s^{(k)}(x,y)$ とは，任意の 2 つのノード x，y の間に k ホップ以内のリンク関係があるかどうかを調べるオペレータであり，次のように定義される．

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

例えば $k=1$ であれば，2 つのノードが存在したときその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．

$t(x,y)$ とは，任意の 2 つのノード x，y 間の距離を求めるオペレータであり，次のように定義する．

$$\begin{aligned} t(x,y) &= \text{distance between } x \text{ to } y \\ &= \arg\min_k \{s^{(k)}(x,y)=1\} \end{aligned}$$

$u_x(y,z)$ は，任意の 2 つのノード y，z の最短パスがノード x を経由するかを判定するオペレータである．これを次のように定義する．

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

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(n 個のノードを持つノード集合) 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ク関係を考えるオペレータ $s^{(1)}$ 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，オペレータ t_x が定義される．

t_x とは，ノード x からノード N に属するそれぞれのノードへの距離を測るオペレータであり☆，次のように定義する．

$$t_x \circ N = \{t(x,k) | k \in N\}$$

ステップ 2 では以上 4 つのオペレータを定義する．

4.3 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．例えば，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．

4.3.1 2 つの値の統合

これら 3 つのステップに分けられたオペレータ以外に 3 ステップにわたるオペレータを適用して得られた 2 つの値の割合を統合するオペレータを考えることができる．なぜなら本稿で対象とするリンクに基づく分類タスクにおいては，分類対象のラベルによる属性値の違いを求めることが有益であると考えられるからである．例えば，ノード x の次数 $Sum \circ t_x \circ C_x^{(1)}$ の値として 5 を，その場合に正のノード集合 N_p による制約を付加した場合 $Sum \circ t_x \circ (C_x^{(1)} \cap N_p)$ の値として 3 を得たとすると，この割合 $Sum \circ t_x \circ (C_x^{(1)} \cap N_p) / Sum \circ t_x \circ C_x^{(1)}$ は $3/5 = 0.6$ として得られる．この値は，ノード x の持つリンクのうちどれだけが正のノード集合に対するリンクであるかを表したものであり，分類問題において重要な属性になりうると考える．これに対応するものとして,「割合」のオペレータ “ratio” を定義する．分類対象のラベルによる属性値の違いは，正のノード集合 N_p での制約があるかないかによる属性値の違いにな

☆ $t_x \circ N$ はノード x とノード集合 N に属するそれぞれのノードとの距離を測るものであり，ノード集合 N に属する任意の 2 つのノード間の距離を求める t とは異なる．

表 2 オペレータリスト

ステージ	Notation	入力	出力	説明	手法
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

る．そこで本稿ではオペレータ “*ratio*” が重要と考え，こうして得られる属性値を $ratio\,of(C_x^{(k)} \cap N_p : C_x^{(k)})$ と表す．この他にも 2 つの値を統合するオペレータは加減乗算などが考えられるが，本稿では分類タスクを対象としているため「割合」のオペレータのみを考える．

ただし，本提案手法ではいくつかのオペレータを無限に生成することができる．例えば，フェーズ 1 において距離に基づくノード集合を考える際，k ホップまでのノード集合は $k = 1 \sim \infty$ まで無限に生成可能である．しかし実際には，距離 k の数を伸ばしても集合に属するノードの数が増えるに従って生成される属性の値は収束するものと考えられる．そのため本稿ではこれを $k = \{1, \infty\}$ のように制限を付加する．フェーズ 2 でも同様に，$s^{(k)}$ をもっともシンプルなオペレータ $s^{(1)}$ のみに制限する．また先に述べたように，割合をとるオペレータを正のノード集合による制限があるかないか，つまり $ratio\,of(C_x^{(k)} \cap N_p : C_x^{(k)})$ に制限する．

本稿で用いるオペレータをまとめたものが，表 2 である．本表よりステップ 1〜3 でそれぞれ 4 つのオペレータを定義している．各ステップでひとつずつオペレータを選択することで，$4 \times 4 \times 4 = 64$ のオペレータの組み合わせができる．さらに割合を考えることで $C_x^{(1)}$ と $N_p \cap C_x^{(1)}$ のノード集合を元に求めた属性値の割合, $C_x^{(\infty)}$ と $N_p \cap C_x^{(\infty)}$ のノード集合を元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が生成され，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[2] 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: $Sum \circ s^{(1)} \circ (C_x^{(1)} \cap N_p)$

上記の社会ネットワーク分析で用いられる属性が生成されるのは，オペレータの設計時にその分析の対象としたためだが，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．例えば，

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

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益な属性になりうる候補であると考えらる．ただし，これらの新たな属性の中には属性として有益性が少ないものもある．例えば，$Max \circ s^{(1)} \circ C_x^{(\infty)}$ 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常に 1 をとるものであり，分類を行う上で有益ではない可能性が高い．さらに生成された新たな属性に対して評価を行ってどの属性が有益な属性かを発見することが必要だと考えられる．そこで以下，本稿における実験では，生成された各属性を評価し新たな有益な属性についても論じる．

5. 実験結果

本章では提案手法の評価を行う．Cora とアットコスメの 2 つのデータセットに対して，本手法を用いて

生成した属性を元に分類を行い，本手法がリンクに基づく分類において有用であることを示す．本実験では，データセット中の各ノードをあらかじめ決められたカテゴリに分類することを考える．また，生成された属性のうちどの属性がこの分類問題において効果的に働いているのかを調べる．

5.1 データセットと実験方法

本章における実験は次の 2 つの評価を行う．

(1) 提案手法がリンクに基づく分類において有益であるか．

(2) 提案手法により生成された属性のうち，どの属性が分類に有益であり，それらの属性のうち有益だがまだ知られていない属性はあるか．

(1) の評価は，Backstrom らの研究[2] での評価手法に基づいて行う．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，これらの属性を元に c4.5 法[12] 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益性を示すため，表 2 に示すように，手法 1 ～ 4 まで段階的にオペレータを増やすこととした．はじめに手法 1 では 1 と書かれたオペレータを用い，手法 2 では 1 と 2，手法 3 では 1 から 3 までを用いるという形をとる．

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 A. McCallum[9] らによって作られたもので，Cora の論文データベースより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たものである．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エッジとする論文ネットワークを構築した．ただし論文間の引用関係は，すべて双方向リンクとして処理した．

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関わらず，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．

表 3 対象とした研究分野
Table 3 Research fields.

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/

5.1.2 アットコスメ データセット

アットコスメとは 100 万人以上のメンバーを持つ，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，感想を書くなどができる．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ーをお気に入りメンバーとして登録することができる．またユーザは様々なコミュニティに所属することができる．これより各ユーザをノードとし，そのお気に入り関係をエッジとした社会ネットワークを構築した．ただしお気に入りリンクは一方向性のリンクであるが，これを双方向リンクとして扱った．

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ー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ーの集合とした．タスクとしては，各ユーザを各コミュニティに分類することを考えた．ただし，コミュニティの選択は，所属メンバー数が 1000 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．

(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，分類に有益な指標である．そこで決定木の上位に現れる属性ほど有益度が高くなるよう，深さ r に現れる属性に $1/r$ の点

表 4 対象としたコミュニティ

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

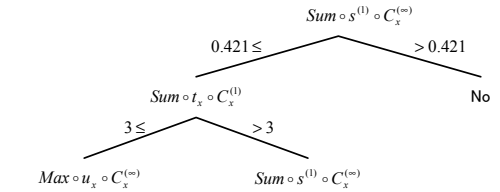


図 2 Cora データセットにおける手法 2 の決定木の深さ 3 までのノード.

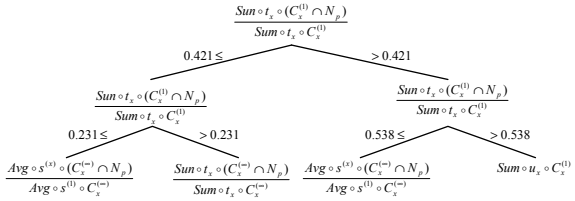


図 3 Cora データセットにおける手法 4 の決定木の深さ 3 までのノード.

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益度として評価した．例えば，Cora のデータセットでは，表 3 に示されたそれぞれのカテゴリの決定木の各ノードに点数をつけ，それらを各属性ごとに足し合わせることで，属性の評価を行う．

5.2 提案手法の有益性の評価

ここでは，リンクに基づく分類タスクに対する提案手法の有益性を評価した結果を示す．ただし，再現率，適合率，F 値の評価は 10 分割交差検定で行った．またその際，生成された決定木の上位についても示す．

表 6 は Cora の論文データセットの「Artificial Intelligence」内の「Machine Learning」の「Neural Networks」の研究分野を対象に実験を行った結果であり，この中には，1682 のノード (論文) があり，そのうちこの研究分野に所属するノード (正例) は 781 件であった．この結果よりオペレータを増やすにつれて，F 値が改善していることがわかる．また，評価を行った際の決定木の上位ノードをみる．図 2 は手法 2 の決定木の深さ 3 までのノードである．図 2 における決定木の最上位ノード $Sum \circ s^{(1)} \circ C_x^{(\infty)}$ は，ノード x から到達可能なノード集合におけるエッジの数であり，意味的にはノード x から到達可能なノード集合に限定した時のネットワーク密度に近い指標である．深さ 2 に現れるノード $Sum \circ s^{(1)} \circ C_x^{(1)}$ 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る．これはノード 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 は手法 4 の決定木の深さ 3 までのノードである．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．つまり，近接するノードのうち特定の分野に属するノードの数が多ければ多いほど，ノード x はそのカテゴリーに属しやすいということを意味している．深さ 3 のノードには $\frac{Avg \circ s^{(1)} \circ (C_x^{(\infty)} \cap N_p)}{Avg \circ s^{(1)} \circ C_x^{(\infty)}}$ 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 5 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 5730 のノード (メンバー) 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 2807 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 値がよくなっていることがわかる．また図 4，図 5 はそれぞれ手法 2，4 の際の決定木の上位ノードである．図 4 における最上位ノード $Sum \circ s^{(1)} \circ C_x^{(1)}$ はノード x の近接ノード集合におけるエッジの数であり，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 になるものである．図 5 における最上位ノード $\frac{Sum \circ to (C_x^{(1)} \cap N_p)}{Sum \circ to C_x^{(1)}}$ 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 $\frac{Sum \circ t_x \circ (C_x^{(\infty)} \cap N_p)}{Sum \circ t_x \circ C_x^{(\infty)}}$ は到達可能なノードの数に対する正のノード数の割合である．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

表 5 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.

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

表 6 Cora のデータセットにおける再現率，適合率，F 値の変化.

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

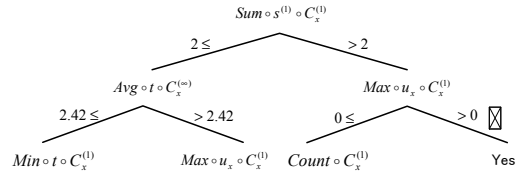


図 4 アットコスメのデータセットにおける手法 2 の決定木の深さ 3 までのノード.

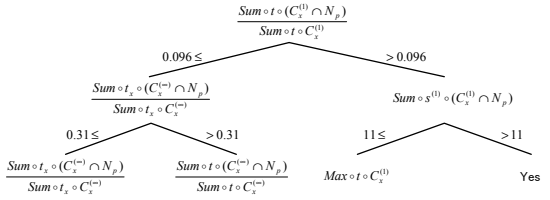


図 5 アットコスメのデータセットにおける手法 4 の決定木の深さ 3 までのノード.

られた「トライアド関係の数」である．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，ひと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 (なぜならすべてのノードは x を介してつながっている)，この指標はノード x のクラスタ係数と近い指標となる．この上位ノードの結果より，すべてのノード集合から得た属性値に対する正のノードに限定した際に得た属性値の割合は多くの場合有益であることがわかる．

5.3 各属性の評価

本節では，各属性の評価を行った結果を示す．

Cora データセットとアットコスメのデータセットでの結果をそれぞれ表 7，表 8 に示す．

この結果より，様々な属性が分類に際して有効であり，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[2] 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わかる．その他にも $Sum \circ s^{(1)} \circ C^{(1)}$ 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となっているが，これらは指標の値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といえる (この例ではクラスタ係数 ($Avg \circ s^{(1)} \circ N_x^{(1)}$) が近い)．これらの結果からわかるように，社会ネットワーク分析で用いられている指標は有益であり，またそれ以外にも社会ネットワーク分析では用いられていない新たな有益な属性があるが，それらの属性の持つ意味は社会ネットワーク分析の指標のそれに近いといえる．

6. 議　　論

本稿で定義したオペレータに加え，新たなオペレータを定義することで，提案手法をさらに拡張することができる．これにより，現手法では得ることができない新たな属性を生成することができる．その例としては，

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

などがある．このほかにも多くのオペレータを考えることができる．例えば，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，中心性の計算を固有ベクトル中心性によって求めることなどが考えられる．ただし固有ベクトル中心性の計算には行列計算が伴うため，オペレータ実装が複雑になり計算コスト困難が問題となる．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益であると結論付けることはできない．そのため新たに様々なオペレータを定義しさらなる分析を進めていくことが今後の課題のひとつである．

また，本稿で提案した手法のパフォーマンスを，ACCA[14] など現在知られている他のリンクに基づく分類アルゴリズムを用いた際のパフォーマンスと比較することで，他の手法に対する本手法の有益性を示すことが必要であると考えている．本稿で提案したアル

表 7 Cora のデータセットにおける有益な上位 10 属性.

Rank	Combination	Description
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[2)].
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有益な上位 10 属性.

Rank	Combination	Description
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.

ゴリズムは *propositionalization* と *upgrade* と呼ばれる帰納論理プログラミング (ILP) において提案されているモデルに含まれると考えられる.

著者らは本稿では, 特にリンクマイニングにおけるリンクに基づく分類タスクに焦点をあてて, 手法を提案し実験を行った. しかし, 適用範囲を広げ提案手法をリンクマイニングにおける他の手法に適用することも考えられる. リンク予測のタスクへの適用を一例として考える. リンク予測とは二つのノード 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. この問題に対処するために次のようなオペレータが求められる.

- 2 つのノード x, y の属性値の集約を行うオペレータ
- 二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

このように提案手法の他のタスクへの適用性を考えることで, 一般に社会ネットワークマイニングのための属性生成を導きたいと考えている.

7. ま と め

本稿では, 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み合わ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することによってノードの分類に有益であることを示した. 2 つのデータセットを用いた実験を通して, 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わかった. 割合という属性は社会学の分野では用いられていないが, この属性も同様に有益である可能性があることが示唆された.

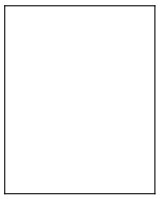
ネットワークと機械学習の分野は, 徐々にその融合領域の研究が進んでおり, 本研究がひとつの重要な知見を提供することになれば, 著者らの幸いとするところである.

参 考 文 献

1) L.Adamic and N.Glance. The political blogosphere and the 2004 u.s. election: Divided they blog. In *LinkKDD-2005*, 2005.
2) L. Backstrom, D. Huttenlocher, X. Lan, and J.Kleinberg. Group formation in large social networks: Membership, growth, and evolution. In *Proc. SIGKDD'06*, 2006.
3) A.-L. Barabási. *LINKED: The New Science of Networks*. Perseus Publishing, Cambride, MA, 2002.
4) A.-L. Barabási. **新ネットワーク思考**. NHK 出版, 2002.
5) L. C. Freeman. Centrality in social net-

works: Conceptual clarification. *Social Networks*, 1:215–239, 1979.
6) N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *Proc. IJCAI-99*, pages 1300–1309, 1999.
7) L.Getoor and C.P. Diehl. Link mining: A survey. *SIGKDD Explorations*, 2(7), 2005.
8) S.Golder and B.A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 2006.
9) A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval Journal*, 3:127–163, 2000. www.research.whizbang.com/data.
10) C. Perlich and F. Provost. Aggregation based feature invention and relational concept classes. In *Proc. KDD 2003*, 2003.
11) A.Popescul and L.Ungar. Statistical relational learning for link prediction. In *IJCAI03 Workshop on Learning Statistical Models from Relational Data*, 2003.
12) J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, California, 1993.
13) J.Scott. *Social Network Analysis: A Handbook (2nd ed.)*. SAGE publications, 2000.
14) P.Sen and L.Getoor. Link-based classification. In *Technical Report CS-TR-4858, University of Maryland*, 2007.
15) S. Wasserman and K. Faust. *Social network analysis. Methods and Applications*. Cambridge University Press, Cambridge, 1994.
16) D.Watts. *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company, 2003.
17) 金. 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.
18) 安. 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.
19) 安. 雪. **実践ネットワーク分析**. 新曜社, 2001.
20) 松. 豊. Web2.0 時代の個人とコラボレーション. **情報処理**, 47(11), 2006.

(平成 17 年 11 月 18 日受付)
(平成 18 年 2 月 4 日採録)



member（正会員）
description