修　士　論　文

社会ネットワークマイニングのための
ネットワーク構造を用いた属性生成に
関する研究

平成20年2月4日提出

指導教員　石塚 満 教授

東京大学大学院 情報理工学系研究科

66412　唐門 準

内容梗概

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．例えば，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，リンクマイニングのタスクであるリンクに基づく分類とリンクの予測を行う．論文，ソーシャルネットワーク，ソーシャルブックマークという３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．

目 次

図 目 次

表 目 次

第1章　序論

1.1　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 [10]．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[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 [16]．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 [31].

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 [32, 30]，例えば，ソーシャルネットワークサービス (SNS) における友人関係 [4, 7]，ブログにおけるコメントやトラックバックを通したユーザの関係 [15]，ソーシャルブックマークにおけるコラボレートタギングを介したタグやユーザの関係 [18, 25, 24]，ウェブ上からの企業間の関係抽出 [35] など社会ネットワークに注目した研究が行われている．特に近年ではウェブが “global village”[34] と称されるように全世界的に大規模に広がるにしたがって，社会ネットワークのマイニング技術の必要性はますます高まっている [40]．社会ネットワークではノードは 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 [39, 36]．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

[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 [16].

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 [38]．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 [33, 9] 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 [8]．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[2]，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，提案手法を，リンクに基づく分類とリンク予測の 2 つのタスクに適用し，論文データベースである Cora のデータセット，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，ソーシャルブックマークであるはてなブックマークのデータセットの 3 つのデータセットを用いて，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．

1 章 序論 本研究の概要について説明する．

2 章 関連研究 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 提案手法 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 評価 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．

[2] 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．

5 章 まとめ 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．

第2章　関連研究

本章では，本研究に関連するいくつかの研究例を紹介する．

本研究では 1 章で述べたように，リンクマイニングにおける 2 つのタスクを対象にネットワーク構造を用いた属性生成手法を提案する．そこでまずリンクマイニングの概要について概説し，本研究で扱うリンクに基づく分類と，リンクの予測の 2 つのタスクについて説明する．その後，ネットワーク構造を用いた属性生成に関する研究例について述べる．

2.1　リンクマイニング

2.1.1　リンクマイニングの概要

リンクマイニングとは前述したように，リンクを解析してそこから有用なデータを取り出すリンク解析，ウェブ上におけるデータマイニング分野であるウェブマイニング [11]，グラフ構造を解析してそこから有用な情報を得るグラフマイニングなどの複合領域であり [16][3]，データマイニングにおけるネットワーク構造の解析分野に位置づけられるものである．

情報は基本的に，独立に分かれているのではなく,，一つ一つのデータの間にはリンク関係が存在し，大規模なデータ構造を持っている．たとえば，WWW では各 Web サイトはリンクを持ち，ほかの Web サイトとリンクを持っている．また学術論文データベースでは他の論文との参照関係を保持している．現在，これらの参照関係に関する研究などが盛んに行われるようになっている．近年では複雑なデータ構造をもつ様々なデータから情報を抽出することなどについての研究が注目を浴びている．

具体的なタスクとしてはノードに関するタスクとネットワーク構造に関するタスクがあり，以下ではそれらについて説明する．

2.1.2　ノードに関するタスク

ノードに関するタスクとしては，下記のようなものがある．以下ではこれらについて説明する．

- ノードのランキング

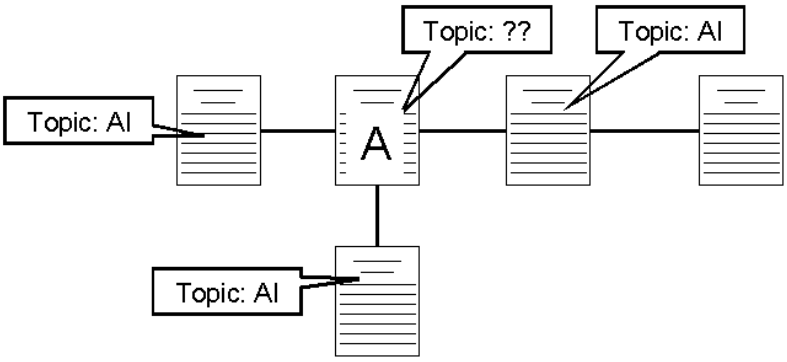


図 2.1: リンクを利用したノード分類の例

- ノードの分類
- ノードのクラスタリング
- ノード解決

ノードのランキング

ノードのランキングとはグラフのリンク構造を利用してグラフ内のオブジェクトの順位付けを行おうとするものである．

この領域に関係する研究として有名なものとして，PageRank がある．Google をはじめとする多くの検索エンジンで使われている PageRank は Web ページをランダムウォークで巡回し，それらに張られているリンクを解析することで，リンクに応じてページの順位付けを行うものである．

ノードの分類

ノードの分類とはグラフ内のオブジェクトのラベル付けを行うものであり，そのラベルとしてとりうる値はあらかじめ与えられている．従来の分類ではオブジェクトの属性などを利用することで分類を行っていたのに対して，グラフのリンク関係を利用することで，リンク関係のあるオブジェクト間には相関が起こりやすいなどの性質を用いることで，分類精度を高めることが行われている．例えば，図 2.1 において，ノード A のトピックがわからないとき，ノード A の近隣のノードのトピックを利用することで，ノード A のトピックが AI ではないかと推測できる．代表的な研究例として，[22, 21] は従来のオブジェクトの属性を用いた分類結果に，各オブジェクトごとに各クラスに対するリンクの分布を考慮するリンクの評価尺度を導入することで，分類の精度を高める研究が行われている．

図 2.2: CiteSeer

ノードのクラスタリング

ノードのクラスタリングとは，グラフのオブジェクトをその特徴が共通するものでグループ化するものである．この分野の代表的な研究例として社会ネットワーク分析 [38] があり，社会ネットワークをネットワーク中の他の人に対するリンクを観測し，同じようなリンク関係を持っている人の集合に分けるといったクラスタ分析の研究が盛んに行われてきた．

また分類に関するタスクと同様に，リンクの分布を考慮して，クラスタリング結果の精度を向上させる研究なども行われている．

ノード解決

オブジェクト同士の関係が与えられているが，実際に存在するエンティティの数がわかっていない場合に，これらの中から同じオブジェクトを特定するものである．

従来の手法では，一般的にオブジェクト属性の類似度を用いて行われていた．この手法を用いた代表例としては，CiteSeer[1] がある．CiteSeer とは図 2.2 のように論文を管理する Web 上のデータベースである．このシステムでは，多くの論文から，引用関係のリストを取り出し，リンク関係をデータベースに格納している．

しかし，このままでは，各エンティティがどれだけ存在しているのかを知ることができず，どのエンティティとエンティティがマッチングするかを考慮することが必要になる．例えば図 2.3 において上と下の文はおそらく同じ論文を表している

Collaborative Interface Agents, Max Metral, and Pattile Maes, Proceedings of the twelfth National Conference on Artificial Intelligence, MIT Press, Cambridge, MA, 1994.

Metral M. Lashkari, Y. and P. Maes. Collaborative interface agents. In conference of the American Association for Artificial Intelligence, Seattle, WA, August 1994.

図 2.3: マッチングすると考えられる引用文の例

と考えられる．そこでこれらのシステムではテキストの類似度などを考えることで，これを解決している．

近年ではさらにリンク各エンティティのリンク関係を考えることで，ノード解決の精度を上げる試みが行われており，例えば [26] においては論文における共参照関係を用いることで，ノード解決の精度を向上するための確率モデルを提案している．

2.1.3 ネットワーク構造に関するタスク

リンクに関するタスクとしてはリンク予測があり，この節ではリンク予測について説明する．

リンク予測

オブジェクトに関するタスクでは，各オブジェクト間のリンクは既出のものであるとしたが，リンク予測では，二つのオブジェクト間にリンクがあるかをそれぞれのオブジェクトの属性値や，ほかのリンクなどを用いて推定するものである．

例えば図 2.4(a) において右のようなグラフが与えられた場合に，この結果を用いてほかのリンクを推定することを考えている．あるいは，図 2.4(b) のようにノードのみが与えられた場合に，そのノードの属性からリンクを推定することが考えられている．

この分野の研究例として [17] は論文の文献関係が既知でないような場合に，リンクの存在に関する確率モデルを構築し，論文の他の属性を用いて論文の引用関係を推定する研究を行っている．

またリンクが一部既知である際に，残りのリンクを推定することで，分類精度の向上につなげようとする研究なども行われている [23]．

このようにリンクマイニングには様々なタスクがあるが，本研究ではこれらのタスクの中でもよく行われる「リンクに基づく分類」と「リンクの予測」の 2 つのタスクについて扱うこととする．

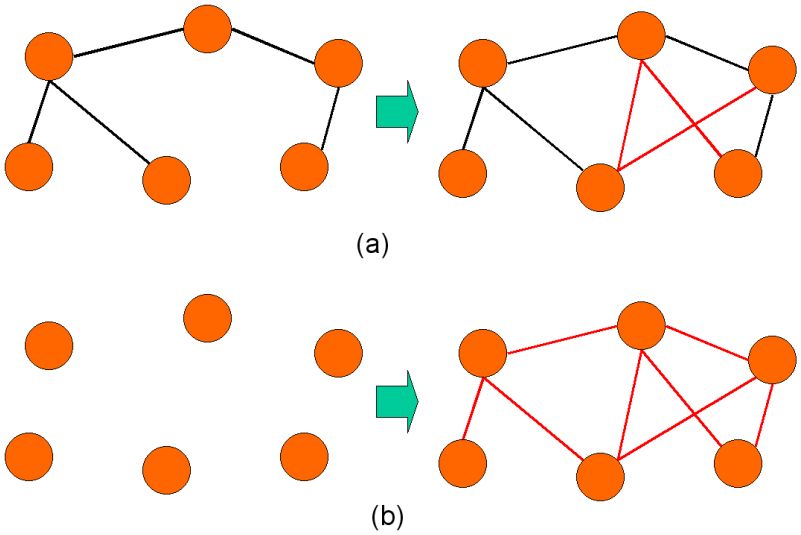


図 2.4: リンク予測

2.2 リンクマイニングの研究例

本節では，前節で挙げたリンクマイニングのタスクを解決するためにどのような研究が行われているかについて概説する．

リンクマイニングにおける研究例では，複雑な関係構造を扱う確率モデルを用いて，ノードの分類や，Citation-Matching，リンク予測などの問題を解決することが必要であり，これらの問題は従来から用いられてきたベイジアンネットワーク [37] のような確率モデルでは，複雑な構造を扱うことが難しい．この問題を解決するために，既存の手法を改良する研究が行われており，その代表例として，Probabilistic Relational Models(PRM)[14] という確率モデルが提案されている．PRM と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ように拡張したものである．そこで本節では，まずその元となるベイジアンネットワークについて説明し，その後，この PRM について述べる．

2.2.1 ベイジアンネットワーク

ベイジアンネットワークとは確率的な因果関係をモデル化したものであり，ノードに確率変数，エッジで確率的な依存関係を表すものである．各ノードはエッジにおいて確率的な依存関係があるノードからは影響を受けるが，それ以外のノードからは影響を受けない，つまり各ノードは親ノードの値がわかっているという条件下で，親ノード以外の各ノードからは独立しているとできる．例えば図 2.5 において，Exam Grade は Good Test Taker と Understands Material を親ノードとして持つので，この 2 つのノードの確率変数にのみ依存し，その他のノードの値

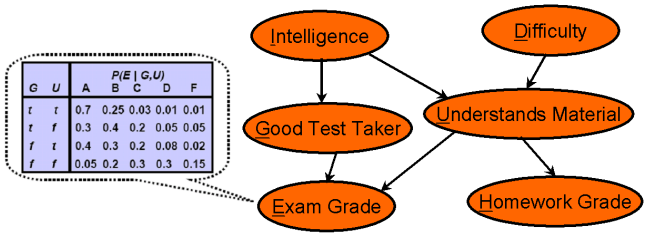


図 2.5: ベイジアンネットワーク

には依存しない．そのため，ある確率変数を X，その確率依存のある親ノードを $Pa(X)$ とおけば，各ノードにおいて条件付確率分布 $P(X|Pa(X))$ が求められる．先の例で，各ノードがある事象をとったとき，モデル全体でその事象が起きる確率は，条件付き確率分布の積

$$P(I,D,G,U,E,H) = \\ P(I)P(D)P(G|I)P(U|I,D)P(E|G,U)P(H|U)$$

で表される．一般に，ベイジアンネットワークにおける各確率変数がそれぞれある値をとったとき，その事象の起きる確率は，

$$P(X_1...X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i)) \tag{2.1}$$

であらわされる．

しかし，ベイジアンネットワークでは各ノードはひとつの属性を持ち，それぞれが独立していると仮定しており，各ノードがその中に様々な属性を持つような場合を扱うことが難しい．またオブジェクトという概念を考えることができないため，同じクラス無いでも属性によって様々なオブジェクトを考えようとするような場合にで起用することができない．そのため複雑な関係構造をもつモデルに対して適応すると関係構造が壊れてしまい，統計的なあいまい性などを招いてしまうという問題点がある．

2.2.2 Probabilistic Relational Models

前項におけるベイジアンネットワークの問題点を解決するためにベイジアンネットワークをより複雑な関係構造に適応できるようにしたのが，PRM である．

PRM の概要

PRM とは図 2.6 のように，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について定義

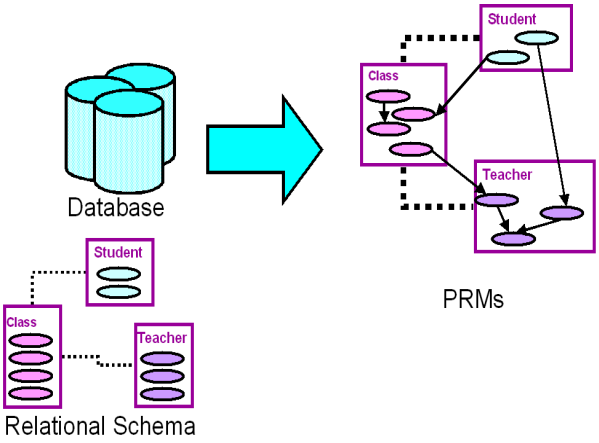


図 2.6: Probabilistic Relational Models

するものである．つまり，PRM とは，リレーショナルスキーマが与えられたとき，それらのすべての可能なインスタンスに対してとりうる値の確率分布を一般的に定義しているものであるといえる．以下ではリレーショナルスキーマと各属性の確率モデルの定義について説明する．

リレーショナルスキーマ

リレーショナルスキーマとは，すべてのクラスとクラスに付随する属性と外部キーの構成さらに外部キーの参照先クラスが与えられたものである．

図 7 の例では，各クラスはエンティティクラスとリレーションシップクラスの 2 つに分類することができる．エンティティクラスとは図 2.7 のおける Paper クラスのように他のクラスを参照していないクラスである．

またリレーションシップクラスとは，Cites クラスのようにクラスとクラスをつなぐいわゆる他のクラスとの関係を表すものである．ここでは，各クラスを $\chi = \{X_1 ... X_n\}$ とする．図 2.7 の例としては，$\chi = \{Paper, Cites\}$ となる．各クラスには，単純な属性値を取る descriptive attribute とほかのクラスを参照する reference slot(データベースにおける外部キー) がある．まず，各クラスの descriptive attribute は，$A\{X\}$ で表し，クラス X における descriptive attribute A は，$A.X$ と表される．また $A.X$ のとりうる値は，$V\{X.A\}$ と表す．例えば図 2.7 において，$A\{Paper\} = \{topic, words\}$，$Paper.Topic \in \{AI, Theory\}$ などと表すことができる．もうひとつの reference slot についても同様に，$R\{X\}$ で表され，クラス X における reference slot ρ は，$X.\rho$ として表される．例として，$R\{Cites\} = \{cited, citing\}$ となる．

また，reference slot ρ には ρ と逆の働きをする，Inverse slot ρ^{-1} を定義することができる．例えば，$Paper.citing_{Cites}{}^{-1}$ で Paper クラスに含まれるインスタンスのすべての引用関係 Cites のインスタンスを返す．

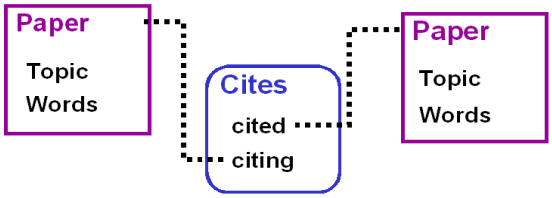


図 2.7: Relatinal Schema

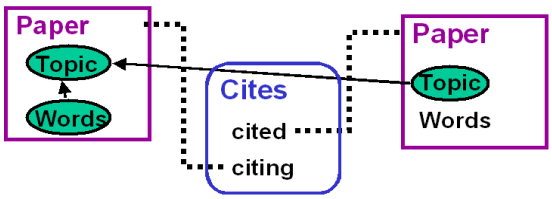


図 2.8: Probabilistic Dependency

また，ρ と ρ^{-1} は連鎖させて slot chain を構成することができる．つまり，$X.\rho_1.....\rho_n.A$ のようなかたちを持ち，クラス同士が直接関係を持たなくとも，間接的に関係を持たせることができる．

各属性の確率モデル

図 2.8 のように各属性に対して，その確率的な依存関係と条件付き確率を定義したものが，確率モデルである．PRM では，ベイジアンネットワークと同様に，リレーショナルスキーマに具体的なインスタンス I が与えられた際にその事象が起こる確率を定義する．

ベイジアンネットワークと同様に各 descriptive attribute に対して，条件付き確率分布が定義される．ただし，reference slot の参照先については既知のものとし，確率モデルには入っていない．

また PRM における各 descriptive attribute に確率的な依存関係のある属性は各クラス内の属性または，reference slot により参照されているクラス，または slot chain による参照先のクラスの属性からとるとする．

これら条件付き確率と確率的な関係構造を学習データにより学習する．

これらの定義により，リレーショナルスキーマに対して，具体的な値 I が与えられたとき，その事象の取りうる確率は，

$$P(I|\sigma_r, \Pi) = \prod_{X} \prod_{x \in O^{\sigma_r}(X)} \prod_{A \in A(x)} P(x.A|Pa(x.A)) \tag{2.2}$$

で表される．ただし，$O^{\sigma_r}(x)$ はクラス X に含まれるすべてのオブジェクトを表す．

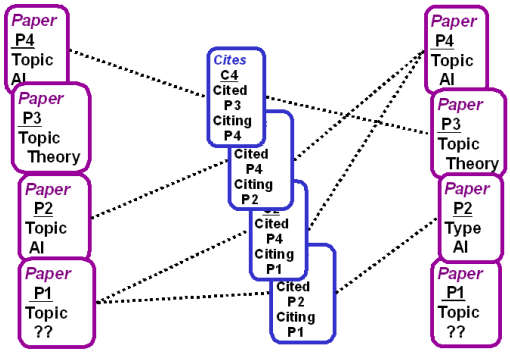


図 2.9: Reference Skeleton

Atrribute Uncertainty

これまで，PRM の定義について紹介したが，本節では実際に PRM がどのように適用されるかについて紹介する．学習された PRM に対して図 2.9 のような実際のインスタンスを表した reference skeletonσ_r(インスタンスとその reference slot の参照先が示されており，descriptive attribute の値が示されていないもの) が与えられたとする．このとき，PRM における，確率的な依存関係を用いて，各 descriptive attribute の値を推論するのが Attribute Uncertainty である．例えば図 2.9 の場合，Paper P1 の属性 Topic の値がわからない場合，図 2.8 のような PRM が学習されていたとすると，このモデルより P1.Topic は P4.Topic，P2.Topic，P1.Word を親に持ち，これらの属性の値から，P1.Topic の値が確率的に推定される．

Link Uncertainty

上記の例では，各オブジェクトの descriptive attribute のみが不明であるとしていた．つまり referece slot の参照先については，既知としている．しかし，この場合，参照先がすべてわかっていない限り，PRM を適用できないと言う問題点がある．そこで [17] では，reference slot に対しても確率モデルを拡張することを提案している．これらは Reference Uncertainty と Existence Uncertainty の二つのタスクに分けられる．以下では，この 2 つのタスクについて説明する．

Reference Uncertainty　Reference Uncertainty とは，オブジェクト自体についてはわかっているが，その関係，つまり，オブジェクト内の reference slot の参照先がわかっていないような状態をさす．図 2.10 のように，論文の参考文献の数はわかっているが，その参考文献の参照先がどのドキュメントになっているのかがわかっていない状態である．

このような場合，与えられるインスタンスは図 2.11 のようなものになる．これを object skeleton σ_o(エンティティと relationship クラスのオブジェクトが与えら

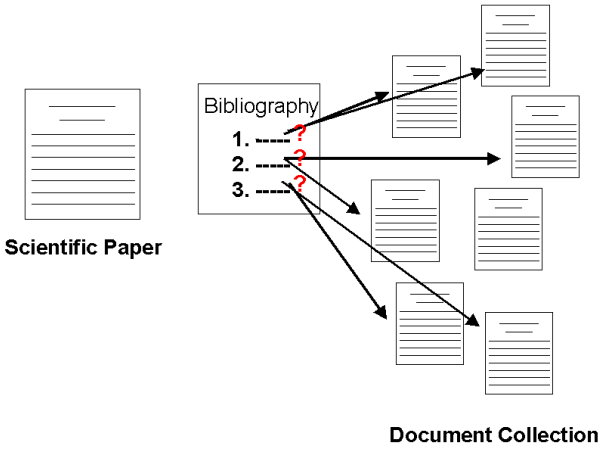


図 2.10: Reference Uncertainty

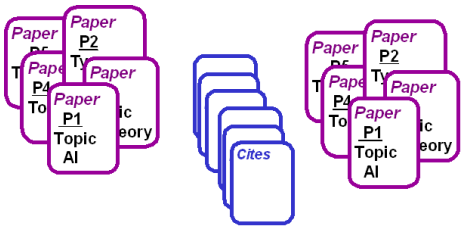


図 2.11: Object Skeleton

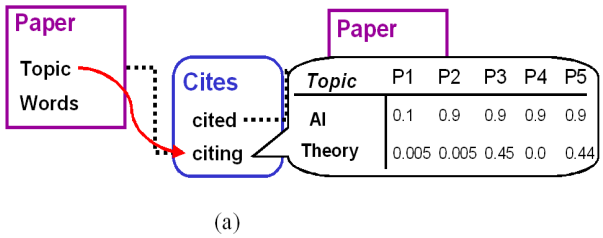
れているが，reference slot の参照先が不明なインスタンスの関係) とする．

このような場合，先のように確率モデルを descriptive attribute のみに作るだけでは，参照先を推測することができないため，reference slot に対しても，確率モデルを作る必要がある．

確率モデルの構築は，descriptive attribute と同様に，参照先の決定に確率的な依存関係のある親属性を学習データで学習し，その値が与えられた下での条件付確率分布を考える．

確率分布を考える対象は，参照先のクラスに属するすべてのオブジェクトすべてに対して考えることになるので (例えば，図 2.11 の例では Paper クラスに属するオブジェクト P1-P5)，図 2.12(a) のような条件付確率分布が学習される．しかし，このモデルでは参照先のオブジェクトの数が増加すると，パラメータの数が膨大になってしまう可能性があるので，一般的な手法として適さない．

そこで，参照先クラスに属するオブジェクトをある属性値に応じてパーティションわけすることを考える．例えば，上記の例では，引用元の論文の Topic が参照先を決定する際に確率的に依存関係があるとしているが，引用元の論文の Topic が AI なら AI に関する論文を引用しやすくなるという推定ができる．そこで，この場合引用先の論文の Topic を元にパーティションわけすることを考える．このパー



(a)

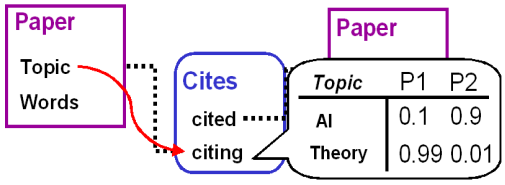


図 2.12: 条件付確率分布

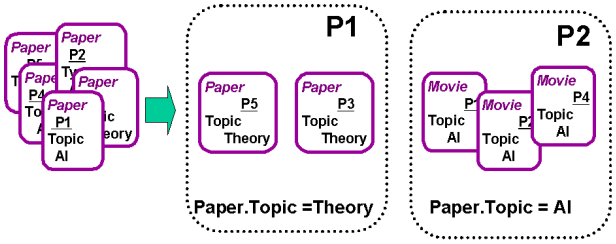


図 2.13: パーティション結果

ティションわけに使う descriptive attribute を partition attribute と呼び，この場合 $partition\ attribute = Paper.topic$ であり，図 2.13 のように分類される．この partition attribute を導入することで先の条件付確率分布は図 2.12(b) のようになり，パラメータの数を減らすことができるようになる．これにより，reference slot はまず partition attribute の値を選び，選択されたパーティション内のオブジェクトをランダムに参照先に選択することで決定される．

このように確率モデルを reference slot まで拡張することで，参照先のわからないインスタンスが存在しても，このモデルを使うことで，参照先を解決することができる．

Existence Uncertainty Reference Uncertainty では，reference slot の参照先は不明であるが，relationship クラスのオブジェクトの数は既知であるとした．Existence Uncertainty では，この relationship クラスのオブジェクトの数も不明であるという条件下での，参照先解決をしようとするものである．例えば図 2.14 のような例では，全論文の組み合わせ 4×3 だけの関係の間に参照関係があるかを解決することを考える必要がある．

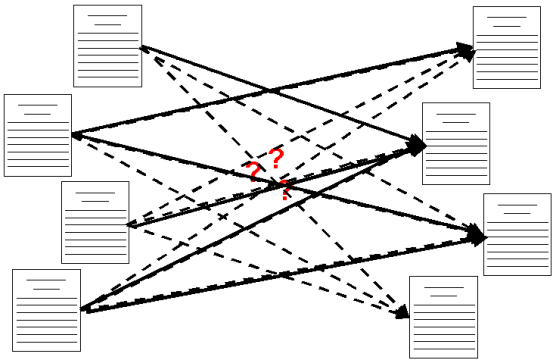


図 2.14: Existence Uncertainty

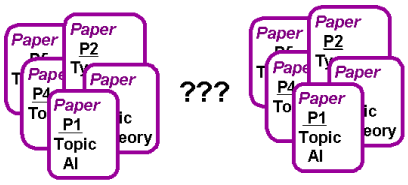


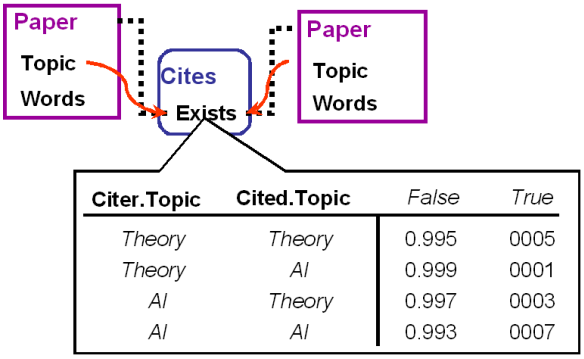
図 2.15: entity skeleton

この場合に与えられるインスタンスの数は図 2.15 のようになり，これを entity skeleton σ_e と呼ぶ (エンティティのみが与えられ，relationship に関してはまったく与えられていない).

この問題を解決するために，図 2.16 のように Exist という仮想的な descriptive attribute を relationship クラスに導入する．この属性は図 2.15 の例でいえば，論文と論文の間に関係があると判断すれば，true をとり，論文と論文の間に関係がないと判断すれば，false の値をとるものである．つまり学習データを与えられた際，図 2.16 のように，Exist に確率的な依存関係のある親属性を決定し，その条件下での条件付確率分布を構築することで，テストデータが与えられた際に，実際に 2 つの論文間に関係が存在するかを推定することができる．

2.3 リンクマイニングの数学的定義

本節では本研究で扱う「リンクに基づくノードの分類」,「リンクの予測」の 2 つのタスクの数学的定義について記す.



Citer.Topic	Cited.Topic	False	True
Theory	Theory	0.995	0005
Theory	AI	0.999	0001
AI	Theory	0.997	0003
AI	AI	0.993	0007

図 2.16: Relational Schema

2.3.1 リンクに基づく分類

リンクに基づく分類タスクとは，2.1.2 で述べたように，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 [31] 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，複結合ネットワーク確率伝播法 (loopy belief propagation) が提案されている．

2.3.2 リンクの予測

リンクの予測タスクは次のように数学的に定義される．まず，時間 t におけるネットワークを G_t とし，時間 t から t' までの間に生成されるリンクからなる G_t のサブグラフを $G[t, t']$ とおく．このとき時間 t_0 におけるネットワーク G_{t_0} を (O,L) とする．時間 t_0，t'_0，t_1，$t'_1(t_0 < t'_0 < t_1 < t'_1)$ が与えられたとき，リンク予測とは時間 t_1 から t'_1 の間にあわられるリンク，つまりサブグラフ $G[t_1, t'_1]$ を予測するものである．またその際学習データとしてサブグラフ $G[t_0, t'_0]$ を用いる．

2.4 ネットワーク構造を用いた属性生成に関する研究

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[8]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[1]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2.1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[28]．従来，SRLの分野では，属性間の関係を学習するPRMの研究[14]がよく知られていた．このモデルでは，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おり，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，あくまでPRM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Popesculらの研究であり，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[27]．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．

[1]ブログサービスがサービスの中心ではあるが，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ができ，ソーシャルネットワーキング機能も持ち合わせている．

表 2.1: Backstrom らの研究 [8] 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 ($\|(u,v)\|u,v \in S \wedge (u,v) \in E_C\|$).
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.

- E_C はコミュニティ C 内のエッジである.

2.5 社会ネットワーク分析で用いられる指標

社会ネットワーク分析とは社会学における学問領域の一分野であり，行為者の属性ではなく，その関係からなるネットワークを分析，記述する方法論である [38, 32]. 本説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 [13]．以下ではそのうち本稿で用いる指標について概説する[2].

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

[2] 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 [41].

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[32, 30]，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．

2.6 リンク予測で用いられる指標

Liben-Nowell らは，リンク予測において用いられる属性についてまとめている[20]．リンク予測のタスクにおいては，ノード x と y の間にリンクがあるかどうかを判定するためにスコア $score(x, y)$ を指標として用いる

このようにリンク予測に用いられる指標のうち有名なものとして，graphic distance，common neighbors，Jaccar's coefficient，Adamic / Adar，preferal attachment などがある[3]．これらの概要を以下に示す．但し以下では，d_{xy} はノード x と y の距離，$\Gamma(x)$ はノード x の隣接ノード集合，$|\Gamma(x)|$ はノード x の隣接ノード集合内のノード数を表す．

graphic distance : d_{xy} ノード x と y の間の距離を用いたスコアで，2つのノード間の距離が小さければ小さいほどそのノード間にリンクができやすいという仮説に基づく．このスコアはリンク予測ではもっとも基本的な方法として知られている．

$$Score(x, y) = d_{xy}$$

common neighbors : $|\Gamma(x) \cap \Gamma(y)|$ このスコアはノード x と y のそれぞれの隣接ノードのうち共通するものの数を用いたものである．より多くの共通する

[3]その他にも PageRank，simRank などのスコアがリンク予測では用いられるが，これらの詳説に関しては [20] を参照されたい．

隣接ノードを持つほど，２つのノード間にリンクが張られやすいという考えである．

$$Score(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

Jaccard's coefficient : $\frac{|\Gamma(x)\cap\Gamma(y)|}{|\Gamma(x)\cup\Gamma(y)|}$ ランダムに x または y の隣接ノード集合から一つのノード n を選んだときに，そのノード n がノード x と y に共通する隣接ノードである確率である．common neighbors を隣接ノードの数で正規化したものとも考えられる．

$$Score(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

Adamic/ Adar : $\sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{log|\Gamma(z)|}$ ノード x, y に共通する隣接ノード集合 $|\Gamma(x)\cap\Gamma(y)|$ に属するノード z を考える．このとき，ノード z の隣接ノード集合の数 $\Gamma(z)$ のログをとりその逆数を $|\Gamma(x) \cap \Gamma(y)|$ に属するすべてのノードに対して考えたものがこれである．

$$Score(x, y) = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{log|\Gamma(z)|}$$

preferential attachment : $|\Gamma(x)| \cdot |\Gamma(y)|$ 新たなエッジがノード x に接続される確率は x の次数に依存するという仮定に基にしている．この仮定に基づけば，ノード x と y の間にリンクが張られる確率は，x，y それぞれの次数の積で表せ，

$$Score(x, y) = |\Gamma(x)| \cdot |\Gamma(y)|$$

となる．

第3章　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 [8] 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 3.1 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[1]．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．

[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，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．

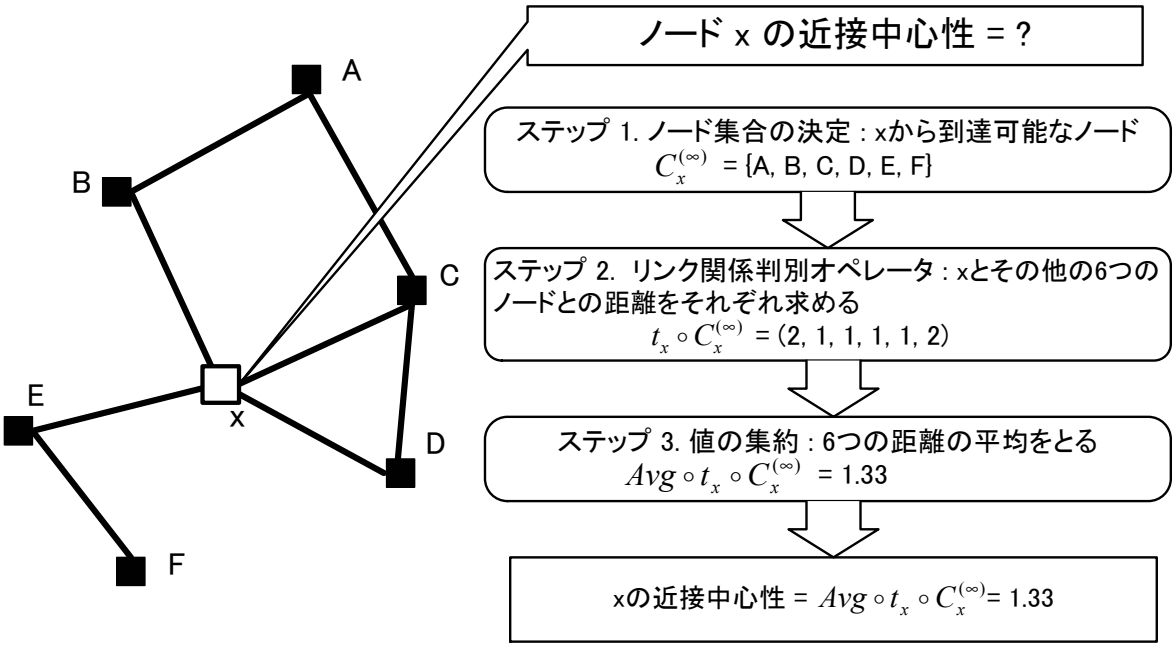


図 3.1: 近接中心性の計算

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.

3.1 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．

3.1.1 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．

$$C^{(k)}(x) = N^{(1)}(x) \cap N^{(2)}(x) \cap \ldots \cap N^{(k)}(x) \tag{3.1}$$

3.1.2 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[2]．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 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．

3.2 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．

[2] 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．

- $s^{(k)}(x,y)$：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[3]，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．

3.3 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．

3.3.1 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．

[3] $t_x \circ N$ 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．

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[4]．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 3.1 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 [8] 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，

[4]ただし，$k = \infty$ に関しては，k を大きくするにしたがって得られる属性値の値は収束する一方計算量は増大するため，$k = 3$ 程度にして実験を行った．

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

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．

3.4 リンク予測への適用

ノード x と y の間のリンクの有無を推定するためには，表 2.6 のようなノード x と y に関するスコア (指標) を属性として生成する必要がある．この属性生成の方法には 2 つの方法がある．一つは，2 つのノード別にそれぞれ属性値を生成し，それをステップ 4 で一つの属性に集約する方法であり，これを**手法 1** とする．例えば，*preferential attachment* ($|\Gamma(x)| \cdot |\Gamma(y)|$) がこれにあたる．ノード x と y のリンクの数をそれぞれ求め，それらを掛け合わせることで求められる．ステップ 4 における 2 つの値を一つにまとめるオペレータとしては，平均を取る Avg，最大値 Max，最小値 Min，差 $Diff$，割合 $ratio$，積 $Product$ の 6 つを定義する．

もう一つの方法は，2 つのノードに注目しネットワーク属性を生成するものであり，これを**手法 2** とする．この手法では，2 つのノードに関連する近隣ノードをノード集合として考える．例えば，*common neighbors* ($|\Gamma(x) \cap \Gamma(y)|$) は，ノード x とノード y の共通する隣接ノードに依存する値である．このようなノード集合を求めるオペレータとして，2 つのノード集合の和と積をとるものを考える．これによりノード x と y 周りのノード集合 $C_x^{(k)}$，$C_y^{(k)}$ をそれぞれ求め，これらのオペレータを適用することにより，新たなノード集合を得ることができる．

さらにより多くのリンク予測のための属性を生成するため，リンクに基づく分類の際とは別に新たなオペレータを各ステップで定義する．まずステップ 2 では，次のようなオペレータ γ を定義する．

$$\gamma(x) = \frac{1}{log|\Gamma(x)|} \quad (|\Gamma(x)| : \textit{ the number of links in node } x)$$

このオペレータはノード x のリンクの数の log を取るものである．

またリンク予測では，ノード x とノード y それぞれによるノード集合の積や和に対してオペレータを適用するため，リンクに基づく分類における t_x と u_x の 2 つのオペレータを次のように再定義する．

$$u_{xy}(z,w) = \begin{cases} 1 & if\ shortest\ path\ between\ z\ and\ w\ includes\ l_{xy} \\ 0 & otherwise \end{cases}$$

但し，l_{xy} はノード x と y 間の距離である．またノード x と y はつながっているものと仮定する．

またオペレータ t_x も次のように t_{xy} として再定義される．

$$t_{xy}(z) = Min\{t(x,z), t(y,z)\}$$

リンク予測において用いられるオペレータをまとめると表 3.2 のようになる．

表 3.1: オペレータリスト

ステップ	表記	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
	$C_x^{(\infty)}$	node x	a nodeset	x から到達可能なノード集合	2
	$N_p \cap C_x^{(1)}$	node x	a nodeset	x の近接ノードのうち正のノード集合	3
	$N_p \cap C_x^{(\infty)}$	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
	t	a nodeset	a list of values	ノードペア間のパス長	1
	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
	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
	Sum	a list of values	a value	和	1
	Min	a list of values	a value	最大値	1
	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

- 各ステップからオペレータを一つずつ取り出して組み合わせる．例：$Avg \circ s^{(1)} \circ C_x^{(k)}$.

表 3.2: オペレータリスト (リンク予測)

ステップ	表記	入力	出力	説明	手法 1	手法 2
1	$C_x^{(k)}$	node x	a node set	x から距離 k までのノード集合	√	
	$C_y^{(k)}$	node y	a node set	y から距離 k までのノード集合	√	
	$C_x^{(k)} \cap C_y^{(k)}$	node x and y	a node set	x から距離 k かつ y から距離 k のノード集合		√
	$C_x^{(k)} \cup C_y^{(k)}$	node x and y	a node set	x から距離 k または y から距離 k のノード集合		√
2	$s^{(k)}$	a node set	a list of values	リンクがあれば 1，それ以外は 0	√	√
	t	a node set	a list of values	ノードペア間のパス長	√	√
	t_x	a node set	a list of values	ノード x とそのほかのノードの距離	√	√
	γ	a node set	a list of values	各ノードの次数		√
	u_{xy}	a node set	a list of values	最短パスが l_{xy} を経由するなら 1，そうでないなら 0	√	√
3	Avg	a list of values	a value	平均	√	√
	Sum	a list of values	a value	和	√	√
	Min	a list of values	a value	最大値	√	√
	Max	a list of values	a value	最小値	√	√
4	$Diff$	two values	value	2 つの値の差	√	√
	Avg	two values	value	2 つの値の平均	√	√
	$Product$	two values	value	2 つの値の積	√	√
	$Ratio$	two values	value	2 つの値の割合	√	√
	Max	two values	value	2 つの値の最大値	√	√
	Min	two values	value	2 つの値の最小値	√	√

- 各ステップからオペレータを一つずつ取り出して組み合わせる．例：$Avg \circ s^{(1)} \circ C_x^{(k)}$.
- 手法 1 におけるステップ 4 のオペレータはオプションである．(はじめに 2 つのノード x と y のノード集合を組み合わせて得られるノード集合を考えており，手法 2 のように 2 つのノードごとに別々に属性を生成し最後にそれを一つにまとめる必要はない．)

第4章　評価

本章では提案手法の評価を行う．Cora，アットコスメ，はてなブックマークの3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．

4.1　データセットと実験方法

本節では提案手法について評価を行った結果について述べる．

4.1.1　実験概要

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 [8] 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3.1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 [29] 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3.1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．

4.1.2 評価に用いる指標

本評価では，再現率，適合率，F 値を用いて評価を行う．各評価値の定義は次のように表される．ただし，あるカテゴリに属しているノードを C，実験の結果そのカテゴリに属すると判定されたノードを N，正と判定されたうち実際にそのカテゴリに属しているノードの数を $R = N \cap C$ とおく．

$$\text{再現率} = \frac{|R|}{|C|}$$

$$\text{適合率} = \frac{|R|}{|N|}$$

$$F \text{ 値} = \frac{2 \times (\text{再現率}) \times (\text{適合率})}{(\text{再現率}) + (\text{適合率})} = \frac{2|R|}{|F| + |C|}$$

4.1.3 データセット

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．

Cora データセット

このデータセットは A. McCallum [23] 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 4.1 のとおりである．

表 4.1: 対象とした研究分野

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/

アットコスメ データセット

アットコスメ[1]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 4.2 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

[1]http://www.cosme.net/

表 4.2: 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

のデータセットでは，表 4.1 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．

はてなブックマークのデータセット

はてな[2]は株式会社はてなが運営する，巨大なコミュニティサイトであり，ブログホスティングサービス「はてなダイアリー」や,「はてなキーワード」「人力検索はてな」などの知識の共有サービス,「はてなフォト」の写真共有サービス等の各サイトの総称である．**はてなブックマーク**[3]とは，ソーシャルブックマーク (collaborative tagging)[4]サービスであり，各ユーザはウェブページにおいて自分のブックマーク (お気に入りリンク) を作成することができ，それぞれのブックマーク (URL) に対して自由にタグや 100 文字以下のショートコメントを残すことができる．また各ユーザは URL やタグを介して他のユーザをリンクされており，他のユーザのブックマークを閲覧することも可能である．本実験で用いるソーシャルブックマークのデータは，2005 年 11 月から 10ヶ月間に作成されたブックマーク約 700 万件を記録したものであり，各ブックマークのデータは「ユーザ名」,「リソース (URL)」,「タグ」，そして「データ (コメント)」から構成されている．

このデータセット内には厳密な意味でのソーシャルネットワークは構築されていないが，URL やタグを介すること社会ネットワークを見出すことができる．本

[2]http://www.hatena.ne.jp/
[3]http://b.hatena.ne.jp/
[4]*del.icio.us*(http://del.icio.us/) が有名である．

表 4.3: 対象としたタグ

タグ名
game
music
software
book
web
news

実験では次のようなURLを介したユーザ間のネットワークを考える．はじめに式4.1のように，ユーザx，y間のURLのオーバーラップの数を考えることで，ユーザ間の類似度を測る[5]．

$$LinkWeight(x,y) = log\frac{|U(x) \cap U(y)|}{|U(x)| \cdot |U(y)|}, \tag{4.1}$$

ただし，$U(x)$ はユーザxがブックマークしたURLの集合であり，$|U(x)|$ はその数である．本実験では式4.1に従い，類似度を全ユーザペアについて求め，類似度が閾値より大きい場合にはノード間にリンクを張ることとしてネットワークを構成した．ただし閾値は，予備実験を行い各ユーザの次数平均がアットコスメにおける次数平均にほぼ等しくなるように，閾値を-4.70に定めた．

実験におけるタスクは，各ユーザが，あるタグを使うか使わないかを推定する，つまりネットワークを用いることによるタグの予測である．学習データとテストデータの作成はCoraやアットコスメのデータセットと同様に行う．データセットは，対象としているタグを使っているノードを正例，対象タグは使っていないがそのタグを使っているユーザにリンクを持っているノードを負例とした．対象タグは1000人以上のユーザが使っているタグのうちランダムに表4.3のタグを選んだ．

4.2 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．

交差検定とは，同じデータを分割して学習データととテストセットを得て評価するときに用いる手法で，一般にk交差検定を行う際には，データをランダムに

[5] この式は相互情報量の式にあたる．

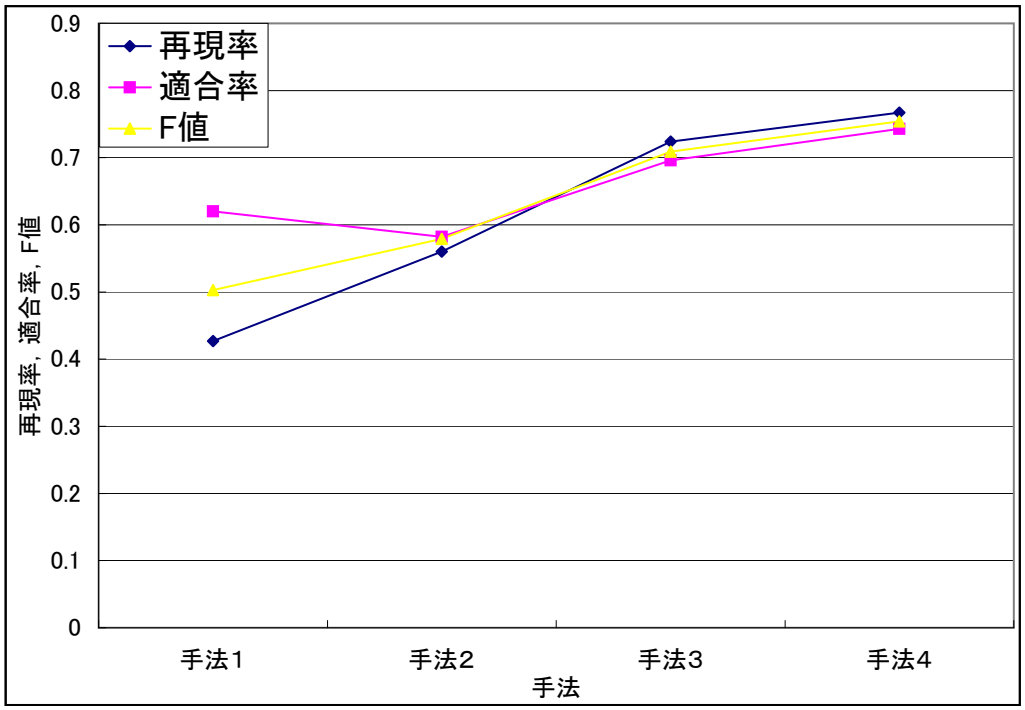


図 4.1: Cora のデータセットにおける再現率，適合率，F 値の変化.

k 等分する．k 等分したデータをそれぞれ，$D_1, D_2, ..., D_k$ とすると，これを

$$\begin{array}{c} (\text{学習データ , テストセット}) \\ (D_2\ D_3 ... D_k\ ,\ D_1) \\ (D_1\ D_3\ D_4 ... D_k\ ,\ D_2) \\ (D_1\ D_2\ D_4\ D_5 ... D_K\ ,\ D3) \\ \vdots \\ (D_1\ D_2 ... D_{k-1}\ ,\ D_k) \end{array} \tag{4.2}$$

とし，これら k 回の評価値の平均をとることで，結果の偏りを防ぐことができるものである．

また評価の際，生成された決定木の上位についても示す．

図 4.1 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 4.2 は手法 2 の決定木の深さ 3 までのノードである．図 4.2 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

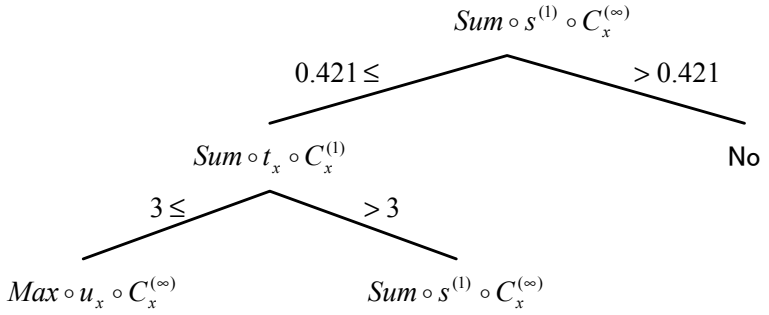


図 4.2: Cora データセットにおける手法 2 の際の決定木の深さ 3 までのノード.

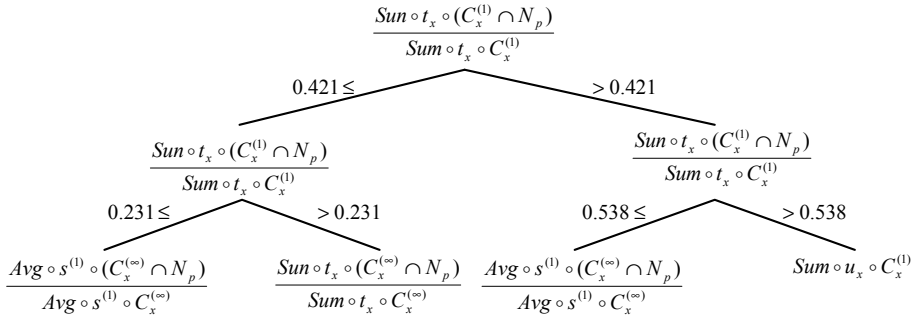


図 4.3: Cora データセットにおける手法 4 の際の決定木の深さ 3 までのノード.

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 のノードに現れる $Max \circ u_x \circ C_x^{(\infty)}$ は，新しい属性である．これはノード x が到達可能なノードのいずれかの最短パス上に存在していれば 1，そうでないとき，0 を取るような値である．

図 4.3 はすべてのオペレータを用いて属性を生成した際の決定木の深さ 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，$Sum \circ u_x \circ C_x^{(1)}$ は，ノード x の近接ノードにおける媒介中心性である．

図 4.4 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．

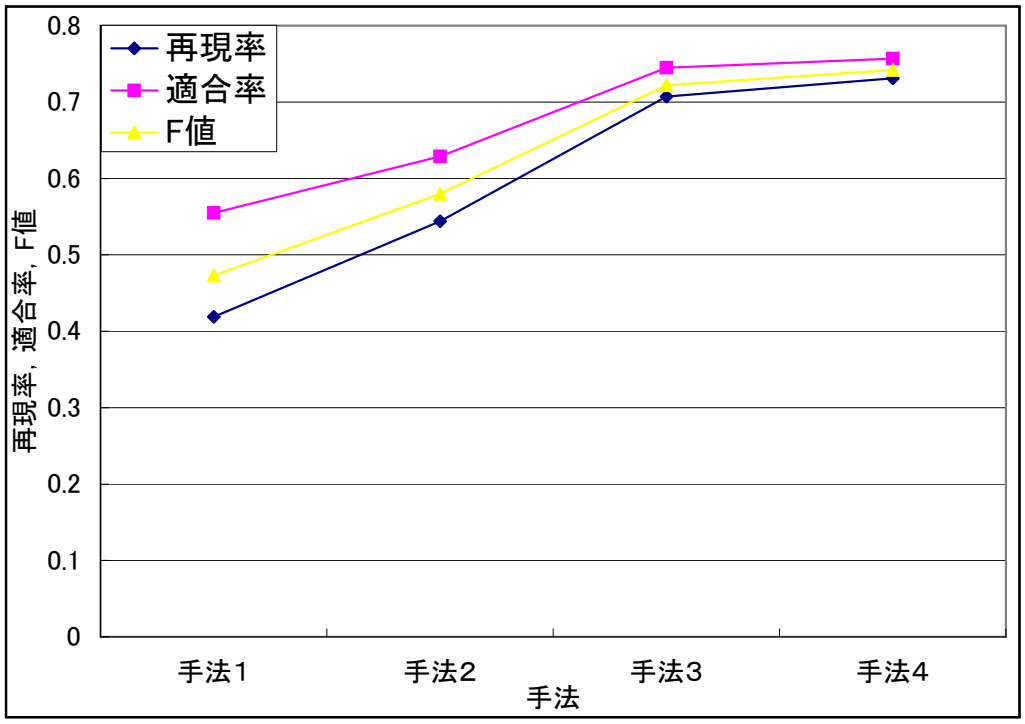


図 4.4: アットコスメのデータセットにおける再現率，適合率，F 値の変化.

また決定木の上位ノードは，タイプ 2 までのオペレータを用いた際は図 4.6，すべてのオペレータを用いた際の結果は図 4.7 のようになった．図 4.6 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 $(Avg \circ s^{(1)} \circ N_x^{(1)})$ 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 4.7 における最上位ノード $\frac{Count \circ (C_x^{(\infty)} \cap N_p)}{Count \circ C_x^{(\infty)}}$ 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．また深さ 2 のノード $Max \circ t \circ (C_x^{(1)} \cap N_p)$ 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る (なぜならすべてのノードは x を介してつながっている)．それゆえ，この指標は先の $Max \circ u_x \circ C_x^{(1)}$ と同様にノード x のクラスタ係数と近い指標となる．

はてなブックマークのデータセットにおける「software」のタグに対して実験を行った結果は図 4.5 のようになった．ただし，このデータ内には正例このうち正例は 1203，負例は 1195 件あった．一般的な傾向はアットコスメのデータセットによる実験と似ており，オペレータを増やすにしたがって結果が改善している．

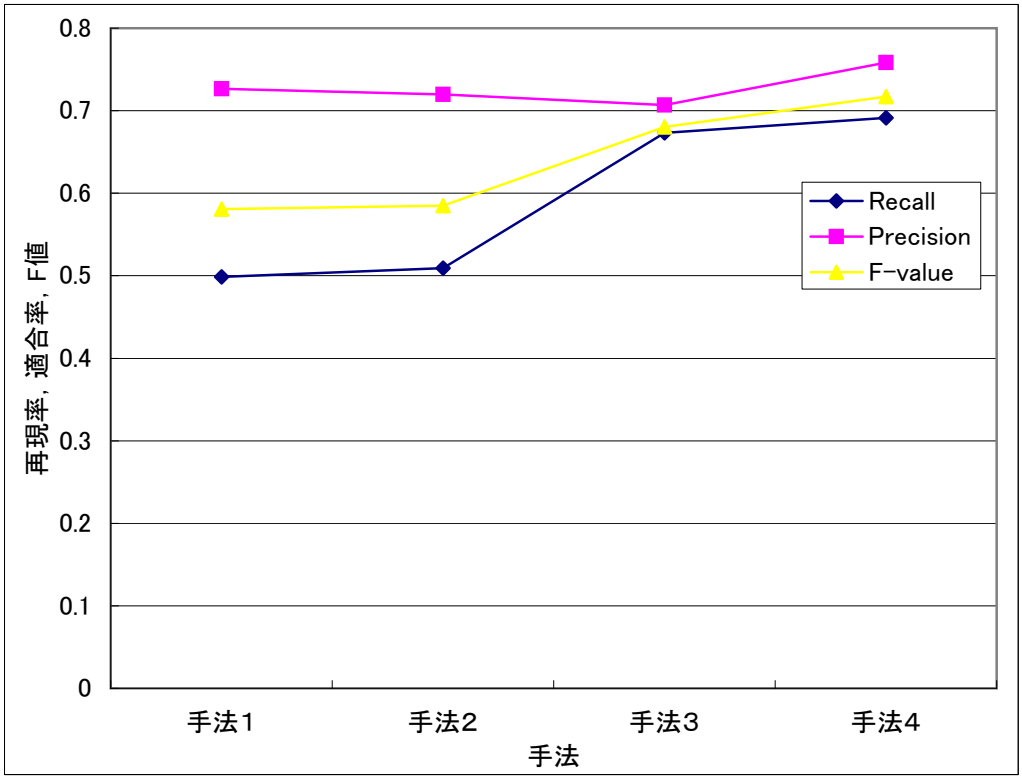


図 4.5: はてなブックマークのデータセットにおける再現率，適合率，F 値の変化.

4.3 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 4.4，表 4.5 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次数 $(Sum \circ t_x \circ C_x^{(1)})$ や，ネットワーク密度 $(Avg \circ s^{(1)} \circ C_x^{(\infty)})$，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 [8] 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という点で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)．また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．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を得ることができる.

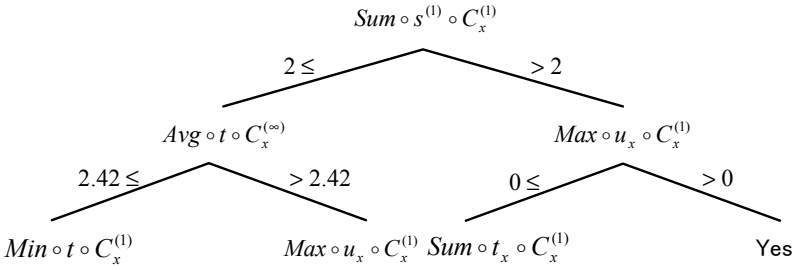


図 4.6: アットコスメのデータセットにおける Stage2 の際の決定木の深さ 3 までのノード.

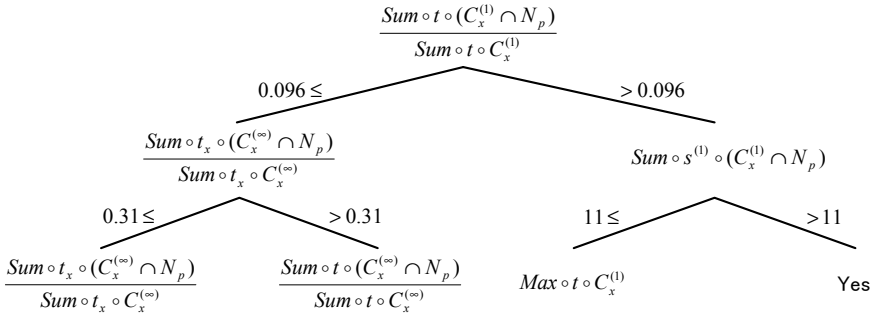


図 4.7: アットコスメのデータセットにおける Stage4 の際の決定木の深さ 3 までのノード.

表 4.6 ははてなブックにおいて実験を行ったときの上位 10 属性である．表の上位にはネットワーク密度や近接中心性などの社会ネットワーク分析での指標がみられる．このような属性は対象としているノードとその周りにあるノードとの間の距離が小さいほど，対象ノードは対象としているタグを使う傾向があったりユーザ自身が興味を持っていることがわかる．この結果はアットコスメのデータセットとは少し異なる結果となった．これはアットコスメのネットワークとはてなブックマークのネットワークではリンクの特徴が異なることが原因と考えられる．アットコスメのネットワークでは 2 つのノードはあるユーザが他のユーザを気に入ったときにリンクされる．それに対して，はてなブックマークのネットワークではリンクは二人のユーザの類似度によってリンクが張られる．つまりアットコスメのリンクはユーザによって自発的に形成されるリンクであり，そのリンクは強いものであるのに対し，はてなブックマークのリンクはユーザ同士で受動的に張られるリンクであるため，リンクの強さは前者に比べて小さくなると考えられる．例えば，表 4.5 における上位 2 番目の属性は友人間のトライアド関係の数であるが，これは近接ノードの重要性を指し示しているものであると考えられる．一方で，最上位属性は対象ノードから到達可能なノード集合を元に得られる属性である．これはノード間の関係が類似度によって構成されるものであるため，その関係は推

表 4.4: 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ネットワーク密度.
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 [8].
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.

移的になりがちであるためだと思われる．提案手法はこのようにその構成された背景や状況が異なるネットワークにも適用可能であり，その点から一般的でありかつ適応性があるといえる．

4.4 リンク予測の評価

リンク予測のタスクに対しては，属性生成手法として**手法 1** と**手法 2** の 2 つの方法を考える．また提案手法との比較として，2.6 節に示された手法を用いて実験を行った．これらの指標はリンク予測においてよく用いられる指標であり，比較手法としては適切であるといえる．

評価は再現率，適合率，F 値を 10 分割交差検定を用いて求めることで行った．図 4.8 はその結果である．結果より手法 1，手法 2 のいずれにおいても F 値は提案手法が他の手法を上回っていることがわかる．特に preferential attachment を除く，common neighbors，Jaccard's coefficient，Adamic/Adar の 3 手法は提案手法に比べて結果が低くなっている．これは，2 つのノードに共通するノード集合がほとんど存在しない，つまりランダムに 2 つのノードを選んだとき，その間に共通する隣接ノードがある確率はほとんどないためである．本研究での提案手法では，単純な共通する隣接ノードだけではなく，距離 2 以上先の共通するノードや隣接ノード集合の和を考慮しているため，結果が向上したものと考えられる．

表 4.5: 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 (リンクに基づく分類タスク).

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.

リンク予測のタスクに関してもどの属性が有益であるかを分類の際と同様に調べた．表 4.7 はリンク予測の手法 1 を用いたときの上位 10 属性を示したものである．手法 1 では，上位 10 属性のほとんどはネットワーク密度や平均パス長，クラスタリング係数など社会ネットワーク分析においてよく用いられる指標である．注目するべき特徴としては，ほとんどの上位属性が (ステップ 4 において) 最大値 (Max) または最小値 (Min) のオペレータを含む形をとっていることである．これは，ノードが他のノードに対してリンクを持つかどうかはリンクの両側のノードの指標に依存するよりは，一つのノードの指標による影響が大きいことがわかる．

表 4.8 は手法 2 における上位 10 属性である．上位 10 属性のうち 8 つがステップ 2 における γ オペレータを含んでいる．γ オペレータは $Adamic/Adar$[5] から導き出されたオペレータであり，リンクの予測タスクにおいて $Adamic/Adar$ は重要な指標であることがわかる．

これら手法 1，手法 2 の結果より，本研究での提案手法が様々なケースにおいて有益であると考えることができる．特に，隣接ノードでなくこれを距離 k 以上 $(k > 1)$ のノード集合を考えることは単純な隣接ノード集合だけを考えた場合と比較して，精度を向上する上で役立つといえる．このように，リンクマイニングにおいては，より広い範囲でノード集合を考えて生成されたの属性は潜在的な重要

表 4.6: はてなブックマークのデータセットにおける有益な上位 10 属性.

Rank	Feature	Description
1	$Max \circ t \circ (C_x^{(\infty)} \cap N_p)$	ノード x から到達可能な正のノード集合における直径.
2	$Max \circ t \circ C_x^{(\infty)}$	ノード x から到達可能なノード集合における直径.
3	$Avg \circ t_x \circ C_x^{(\infty)}$	ノード x から到達可能なノード集合における近接中心性.
4	$Avg \circ t_x \circ (C_x^{(\infty)} \cap N_p)$	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.
5	$Sum \circ t_x \circ C_x^{(1)}$	ノード x の近接ノードの数.
6	$Avg \circ s^{(1)} \circ (C_x^{(\infty)} \cap N_p)$	ノード x から到達可能なノード集合内でのネットワーク密度.
7	$Avg \circ u_x \circ C_x^{(\infty)}$	ノード x 到達可能なノード集合における媒介中心性.
8	$Sum \circ s^{(1)} \circ C_x^{(\infty)}$	ノード x から到達可能な正のノード内でのリンク数.
9	$Sum \circ u_x \circ C_x^{(1)}$	ノード x に隣接するノード集合における媒介中心性.
10	$Max \circ t \circ C_x^{(\infty)} \cap N_p$	ノード x から到達可能な正のノード集合内の最短パスが x を経由していれば 1，そうでなければ 0.

性を持つことを示している.

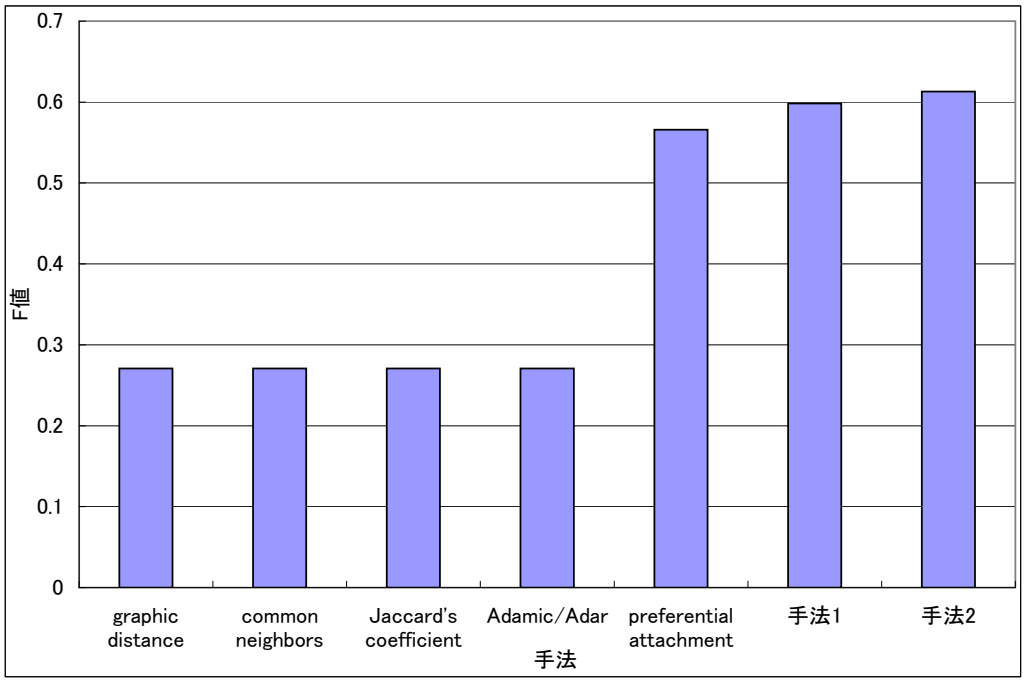


図 4.8: アットコスメのデータセットにおける再現率，適合率，F 値 (リンク予測).

表 4.7: 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 (リンク予測タスク (手法 1)).

Rank	Feature	Description
1	$Max\{Avg \circ t \circ C_x^{(2)}, Avg \circ t \circ C_y^{(2)}\}$	x,y のノード集合における平均パス長の最大値.
2	$Max\{Sum \circ s^{(1)} \circ C_x^{(1)}, Sum \circ s^{(1)} \circ C_y^{(1)}\}$	x,y のクラスタ係数の最大値.
3	$Min\{Sum \circ t_x \circ C_x^{(1)}, Sum \circ t_x \circ C_y^{(1)}\}$	x,y のノードの次数の最小値.
4	$Max\{Avg \circ s^{(1)} \circ C_x^{(2)}, Avg \circ s^{(1)} \circ C_x^{(2)}\}$	x,y のノード集合におけるネットワーク密度の最大値.
5	$Max\{Avg \circ u_x \circ C_x^{(2)}, Avg \circ u_x \circ C_y^{(2)}\}$	x,y の媒介中心性の最大値.
6	$Min\{Avg \circ t \circ C_x^{(2)}, Avg \circ t \circ C_y^{(2)}\}$	x,y のノード集合における平均パス長の最小値.
7	$Max\{Sum \circ u_x \circ C_x^{(2)}, Sum \circ u_x \circ C_y^{(2)}\}$	x,y の媒介中心性における最大値.
8	$Max\{Sum \circ t_x \circ C_x^{(1)}, Sum \circ t_x \circ C_y^{(1)}\}$	x,y のノードの次数の最大値.
9	$Avg\{Sum \circ s^{(1)} \circ C_x^{(1)}, Sum \circ s^{(1)} \circ C_y^{(1)}\}$	x,y のクラスタ係数の平均.
10	$Sum \circ u_x \circ C_x^{(2)} - Sum \circ u_x \circ C_y^{(2)}$	x,y の媒介中心性の差.

表 4.8: 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 (リンク予測タスク (手法 2)).

Rank	Feature	Description
1	$Min \circ \gamma \circ (C_x^{(2)} \cup C_y^{(2)})$	x または y のノード集合におけるノードの次数の最小値.
2	$Max \circ \gamma \circ (C_x^{(2)} \cup C_y^{(2)})$	x または y のノード集合におけるノードの次数の最大値.
3	$Avg \circ \gamma \circ (C_x^{(2)} \cup C_y^{(2)})$	x または y のノード集合におけるノードの次数の平均.
4	$Sum \circ \gamma \circ (C_x^{(1)} \cup C_x^{(1)})$	x または y のノード集合におけるノードの次数の和.
5	$Max\{Max \circ \gamma \circ (C_x^{(1)} \cap C_x^{(1)}),$ $Max \circ \gamma \circ (C_x^{(1)} \cup C_x^{(1)})\}$	x または y のノード集合におけるノードの次数の最大値と x かつ y のノード集合におけるノードの次数の最大値の最大値.
6	$Min \circ \gamma \circ (C_x^{(2)} \cup C_y^{(2)})$	x または y のノード集合におけるノードの次数の最小値.
7	$(Min \circ \gamma \circ (C_x^{(2)} \cap C_y^{(2)})) \cdot (Min \circ \gamma \circ (C_x^{(2)} \cup C_y^{(2)}))$	x または y のノード集合におけるノードの次数の最小値と x かつ y のノード集合におけるノードの次数の最小値の積.
8	$Max\{Min \circ \gamma \circ (C_x^{(2)} \cap C_y^{(2)}),$ $Min \circ \gamma \circ (C_x^{(2)} \cup C_y^{(2)})$	x または y のノード集合におけるノードの次数の最小値と x かつ y のノード集合におけるノードの次数の最小値の最大値.
9	$Sum \circ t_x \circ (C_x^{(1)} \cup C_y^{(1)})$	x または y のノード集合に属するノードの数.
10	$Sum \circ s^{(1)} \circ (C_x^{(1)} \cup C_y^{(1)})$	x または y のノード集合におけるクラスタ係数.

第5章　まとめと今後の課題

5.1　議論

本節では本研究から得られた結果を踏まえ今後の課題について議論する．

5.1.1　新たなオペレータの追加

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[31] 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．

また本研究では属性の生成に主眼をおき，生成した属性に関しては基本的にすべてを用いて各タスクを処理した．しかし今後オペレータを増加させるに従い生成される属性も増大するため，決定木の学習などで過学習状態になることが考えられる．そのため属性選択の処理についての考慮が今後求められる．

5.1.2 本研究の応用性

本研究で対象としたリンクマイニングにおける 2 つのタスク「リンクに基づく分類」と「リンクの予測」の実データに対する応用について考察する．

推薦システムとしてのリンクに基づく分類 SNS などにおいてユーザにあった適切な商品を推奨する，場所や広告を提示する，などのサービスを考えたいとする．このようなシステムは，各ユーザを「ある商品を推薦するのに適切なユーザであるかどうか」,「ある場所を提示するのに適切なユーザかどうか」のように正か負かの 2 値分類タスクとしてみなすことができる．いわゆる推薦システムを本研究でのネットワーク構造を用いて生成することで属性だけを用いた場合に比べてより正確な推薦ができると考えられる [12, 6]．

リンク推薦システムとしてのリンク予測 SNS における友人の推薦，ブログにおいて他のブログへのリンクの推薦，ウィキペディアにおけるほかの単語へのリンクの提示などのいわゆるリンク推薦システムは，現在のネットワーク関係から未来のネットワークを予測するリンク予測のタスクとして扱うことができる．

このようにリンクマイニングにおけるタスクと SNS やブログなど近年 Web2.0[2] と総称されるサービスには強いつながりを見ることができる．SNS の盛り上がりやウィキペディアをはじめとする新たな Web2.0 的サービスが増えていることを考慮すれば，このようなタスクを処理することの重要性はこれからますます高まると考えられる．

5.1.3 ノードの属性値とネットワークの関係

本研究では「ネットワーク構造を用いた」属性の生成を対象としており，分類タスクにおいてノードのカテゴリ属性を用いた以外はノードの属性を用いることはなかった．しかし，実際には，各ノードの分類タスクでは各ノードの属性を用いることがベースにある．例えば，Jensen らの研究では，PRM においてノード自身の属性と隣接の属性の関係性について論じている [19]．Jensen らの研究によれば PRM のモデルは，

- Intrinsic：ノード自身の属性だけ使ってクラスを予測
- R1：隣接するノードの属性を用いる
- R2：距離 2 までのノードの属性を用いる
- CI：隣接するノードのクラス情報を用いる

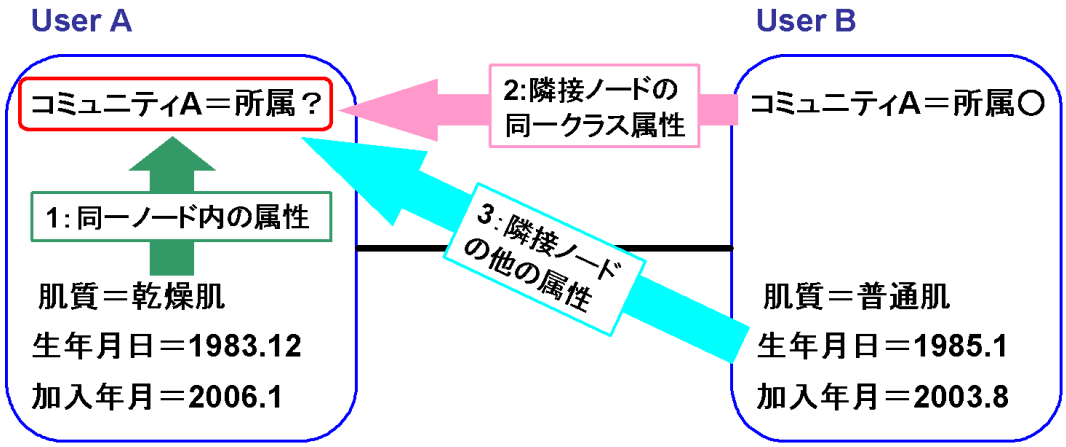


図 5.1: 属性とネットワークの関係

- RCI1：隣接するノードの属性とクラス情報を使う

のような段階にわけられるという．つまり，分類問題では，まずベースとして，そのノードに属性を用いた分類を行い，その次に関係性 (隣接ノード) を用い，最後にネットワーク構造 (距離 2 以上のノード情報) を用いるという流れがある．またクラス同定問題では隣接ノードの属性を用いる際にも，隣接ノードのクラス情報だけでなく，その他の属性が分類に際して重要になるといえる．

この流れを本研究とあわせて考えると本研究の今後の課題として考えられるのはその属性と関係性を融合して用いることである．ではノードの属性とその関係性を用いるにはどのような処理が必要になるかを考える．先に挙げた研究例などを踏まえると，ノードの属性と関係性の間に成り立つ段階を次のように考えることができる．

1:ノードの属性 ノード自身に固有の属性に着目するものである．(縦の関係)

2：隣接ノードのカテゴリ属性 ノードとリンク関係を持つノードの同じクラスの属性に注目する．(横の関係)

3：隣接ノードの他属性 ノードとリンク関係を持つノードのほかのクラスの属性に注目する．(斜めの関係)

このような概念をまとめると図 5.1 のように表せる

例えばアットコスメのデータを例として用いる．各ユーザを様々なコミュニティに属するかどうかを決定するような分類タスクを扱う際，上記の階層構造は次のように考えることができる．

1:ノードの属性 各ユーザ (ノードの) 肌質，誕生日，加入年月などユーザ固有の属性値を用いることで分類を行う．

2：隣接ノードのカテゴリ属性 各ノードと隣接するノードに注目しそのノードのカテゴリ情報に着目することで，分類を行う．例えば隣接するユーザのカテゴリ情報はユーザのコミュニティへの加入に与える影響があると考えられる．

3：隣接ノードの他属性 隣接するノードのカテゴリ以外の属性を用いる．但し，これらの属性の影響はカテゴリ属性よりは低いと考えられる．これらの属性が階層 1 で同じ属性の値が有効に作用している条件で，効いてくる属性であると考えられる．例えば，カテゴリ属性の決定に隣接ノードの肌質の属性値が有効であるかは，階層 1 でカテゴリ属性を決定すべきノードの肌質がカテゴリ属性の決定に有益に働いているという条件に依存すると思われる．

このように各ノードが持つ様々な属性の値とそのノードが織り成すネットワークは関係を持っており，分析においてもこれらを用いることは重要な鍵になると考えられる．

5.2 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．

謝辞

本研究を行うにあたり，非常に多くの方々のご指導・ご鞭撻を賜りましたことを，この場をお借りして心より感謝いたします.

指導教員の石塚満教授には，大変お忙しい中貴重な時間を割いていただき，日頃から多くのご指導・ご鞭撻を賜り，また本研究について指針を与えていただき，心より感謝いたします.

石塚研究室秘書の藤田メイコさんには，平時より楽しく快適に研究を行う環境を整えていただき感謝しています.

土肥浩助教には，ミーティングなどで，ご指導やアドバイスを賜りましたこと，また，本業の傍ら研究室のサーバーやネットワークの管理からマシンの維持まで行っていただき，快適な研究環境の保持に勤めてくださったことを心より感謝いたします.

石塚研究室の OB であり，現在，東京大学大学院 工学系研究科 総合研究機構所属であり松尾ぐみのくみ長である松尾豊准教授には，お忙しい中，毎週研究室に来ていただき，個々の研究へのアドバイスから，研究に対する心構えまで色々とご指導くださって感謝しております．また，単なる研究への指針・アドバイスにとどまらず，論文の手直しや書き方をご教授いただきました．また，ご指導いただいたおかげで，国際学会での発表という貴重な経験ができたことを大変感謝しております．こうしたご指導を通じて，研究というものがどうあるべきかについて深く学ぶ機会を得られたことを深く感謝いたします.

石塚研究室 OB であり，現在辻井研究室で特別研究員をされている岡崎直観氏には有益な指摘やディスカッションをしていただきました．また研究室内でのミーティングを取り仕切ってくださったおかげで，スムーズに研究を進めることができ，感謝しております.

成蹊大学の山本晶先生にはアットコスメのデータを頂き，よい実験評価を行うことができたことを感謝いたします.

石塚研究室 OB であり，現在ドイツ人工知能研究所におられる森純一郎氏には，有益な指摘やディスカッションをしていただきました．また研究室のネットワークの管理でも大変お世話になり感謝しております.

博士 2 年の金英子氏，古川忠延氏，博士 1 年のボッレーガラ ダヌシカ氏には普段のミーティング等で鋭いご指摘を頂き，研究をよりよいものにすることができ，感謝しております．古川氏にははてなブックマークのデータを頂いたこともあわせて感謝いたします.

同じ学年である，岡嶋譲氏，蓑津真一郎氏とはお互いに刺激しあい切磋琢磨して研究を進める一方，同じ趣味をもつ仲間として研究の間の雑談などを通して息抜きの場を得られたことで楽しく研究室での生活を送れたことを感謝いたします．

また，日頃から多岐に渡りご指導くださったその他の石塚研究室，松尾ぐみの先輩・後輩諸氏にも感謝いたします．

そしてなにより温かい目で見守ってくださいました両親に感謝いたします．

最後に，修士課程をこの石塚研究室で過ごせたことは，自分にとって非常に幸運なことだったと思います．2 年間，非常に良い方々に囲まれ，有意義な研究生活を送れたことは今後の人生に必ずや生きるものであると確信しております．皆様本当にありがとうございました．

参考文献

[1] http://citeseer.ist.psu.edu/.

[2] http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

[3] Link mining: A new data mining challenge. *SIGKDD Explorations*, 5(1), 2003.

[4] Lada Adamic and Natalie Glance. The political blogosphere and the 2004 u.s. election: Divided they blog. In *LinkKDD-2005*, 2005.

[5] Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[6] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[7] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Proc. WWW2007*, 2007.

[8] L. Backstrom, D. Huttenlocher, X. Lan, and J. Kleinberg. Group formation in large social networks: Membership, growth, and evolution. In *Proc. SIGKDD'06*, 2006.

[9] Albert-Lászió Barabási. *LINKED: The New Science of Networks*. Perseus Publishing, Cambride, MA, 2002.

[10] Albert-Lászió Barabási. **新ネットワーク思考**. NHK 出版, 2002.

[11] George Chang and Marcus Healey. *Web* **マイニング**. 共立出版, 2004.

[12] A. Felfernig, G. Friedrich, and L. Schmidt-Thieme. Recommender systems. 2007.

[13] L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1:215–239, 1979.

[14] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *Proc. IJCAI-99*, pages 1300–1309, 1999.

[15] T. Furukawa, Y. Matsuo, I. Ohmukai, K. Uchiyama, and M. Ishizuka. Social networks and reading behavior in the blogosphere. In *Proc. ICWSM 2007*, 2007.

[16] L. Getoor and C. P. Diehl. Link mining: A survey. *SIGKDD Explorations*, 2(7), 2005.

[17] Lise Getoor, Nir Friedman, Daphne Koller, and Benjamin Taskar. Learning probabilistic relational models with link structure. *Journal of Machine Learning Research*, 2002.

[18] S. Golder and B. A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 2006.

[19] David Jensen, Jennifer Neville, and Brian Gallagher. Why collective inference improves relational classification. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.

[20] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proc. CIKM*, pages 556–559, 2003.

[21] Qing Lu and Lise Getoor. Link-based classification. *International Conference on Machine Learning*, (Washington, DC), 2003.

[22] Qing Lu and Lise Getoor. Link-based text classification. *IJCAI Workshop on Text Mining and Link Analysis*, MX(Acapulco), 2003.

[23] 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.

[24] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *Proc. ISWC2005*, 2005.

[25] P. Mika. Web semantics: Science services and agents on the world wide web. pages 5–15, 2007.

[26] Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart Russell, and Ilya Shpitser. Identity uncertainty and citation matching. *NIPS*, 2002.

[27] C. Perlich and F. Provost. Aggregation based feature invention and relational concept classes. In *Proc. KDD 2003*, 2003.

[28] A. Popescul and L. Ungar. Statistical relational learning for link prediction. In *IJCAI03 Workshop on Learning Statistical Models from Relational Data*, 2003.

[29] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, California, 1993.

[30] John Scott. *Social Network Analysis: A Handbook (2nd ed.)*. SAGE publications, 2000.

[31] P. Sen and L. Getoor. Link-based classification. Technical Report CS-TR-4858, University of Maryland, 2007.

[32] S. Wasserman and K. Faust. *Social network analysis. Methods and Applications*. Cambridge University Press, Cambridge, 1994.

[33] D. Watts. *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company, 2003.

[34] Barry Wellman. The global village: Internet and community. *The Arts & Science Review, University of Toronto*, 1(1):26–30, 2006.

[35] 金 英子, 松尾 豊, and 石塚 満. Web 上の情報を用いた企業間関係の抽出. **人工知能学会論文誌**, 2007.

[36] 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.

[37] 繁桝 算男. 本村 陽一. 植野 真臣. **ベイジアンネットワーク概説**. 培風社, 2006.

[38] 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.

[39] 安田 雪. **実践ネットワーク分析**. 新曜社, 2001.

[40] 松尾 豊. Web2.0 時代の個人とコラボレーション. **情報処理**, 47(11), 2006.

[41] 松尾 豊 and 松村 真宏. Web コンピューティング. **人工知能学会誌**, 22(4), 2007.

発表文献

- 唐門 準，松尾 豊，石塚 満，“論文ネットワークからのリンクマイニング”，電子情報通信学会信学技報，(2007.1).
- 唐門 準，松尾 豊，石塚 満，“エンティティのネットワーク構造を用いた属性生成”，人工知能学会全国大会，(2007.6).
- Jun Karamon, Yutaka Matsuo, Mitsuru Ishizuka: “Generating Social Network Features for Link-based Classification”, European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Warsaw, Poland, Sep. 2007.
- 唐門 準，松尾 豊，石塚 満，“社会ネットワークマイニングのためのネットワーク構造を用いた属性生成”，情報処理学会全国大会，(2008.3 発表予定).