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

唐門準† 松尾豊‡ 石塚満†

†東京大学大学院情報理工学系研究科
‡東京大学大学院工学系研究科

1 はじめに

近年，Web の発展により SNS やソーシャルブックマークなど，ネットワーク構造を持つデータが多く存在しており，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．例えば Backstrom らの研究 [1] では，ネットワーク構造を用いた有益な属性を発見している．このようなネットワークに着目したデータマイニングは一般にリンクマイニングと呼ばれる．

一方，これまで社会ネットワーク分析 [3] や複雑ネットワークの分野ではネットワークを評価する指標として，中心性，構造空隙，クラスタ係数などが用いられてきた．

本稿では，この 2 つの研究の流れに注目し，社会ネットワーク分析で用いられる指標をはじめ，従来から用いられてきたネットワーク構造を用いた属性の生成を可能とするオペレータを定義し，リンクに基づく分類に適用する．またアットコスメのデータに適用し，提案手法の有用性を確認する．

2 提案手法

本章では，ネットワーク構造を用いた属性を体系的に生成するための手法を提案する．まず社会ネットワーク分析で用いられている様々な指標を分析することで，属性の生成を次の 3 つのステップに分解する．

ステップ 1 対象ノードを決定する．

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

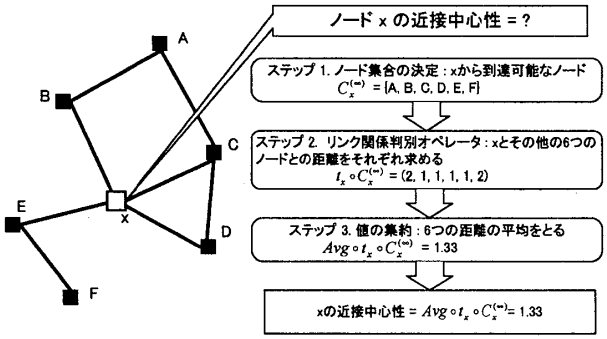

図 1: 属性生成の流れ

ステップ 3 ステップ 2 の結果を集計し属性値を得る．

各ステップでは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，これらの指標をオペレータの組み合わせで実現することを考える．

たとえば，近接中心性[1]の生成は図 1 のように考えることができる．まずステップ 1 として，中心性を求める対象ノード x から到達可能なノード集合を求める．ステップ 2 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．最後にステップ 3 として，得られた各距離を平均することで求める値が得られる．ステップ 1 からステップ 3 までの操作を行うオペレータをそれぞれ，$C_x^{(\infty)}$，t_x，Avg とおけば，ノード x における近接中心性は $Avg \circ t_x \circ C_x^{(\infty)}$ という 3 つのオペレータの組み合わせとして表現できる．

このように，3 つのステップのオペレータを組み合わせることで，さまざまな指標が得られる．

本稿で用いるオペレータをまとめたものが，表 1 である．ステップ 1~3 に対して，それぞれ 4 つのオペレータを定義している．各ステップで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

[1]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ある．

Generating Social Network Features for Mining Social Networks

Jun KARAMON†, Yutaka MATSUO†, Mitsuru ISHIZUKA†
†Graduate School of Information Science and Technology , The University of Tokyo
Bunkyo-ku, Tokyo 113-8656, Japan
‡Graduate School of Engineering , The University of Tokyo
Bunkyo-ku, Tokyo 113-8656, Japan

*

表 1: オペレータリスト

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

- 正のノード集合 N_p とはカテゴリ属性が目的とする属性値をとるノード集合のことである．
- リンクに基づく分類では，すべてのノード集合での値と正のノード集合での値の割合が重要と考えられ，ステップ 4 としてこれらの割合をとるオペレータ $ratio_p$ を付加的に用意する．

$C_x^{(1)}$ 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．

3 実験結果

本章では，定義したオペレータにより生成された属性がリンクに基づく分類タスクに対して有益であるかについて評価を行った結果について述べる．

評価は，アットコスメ[2]のデータセットを用い次のように行った．まず各ユーザをノードとし，お気に入り関係をリンク（方向なし）とした社会ネットワークを構築する．次に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 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 [2] を用いて決定木を学習し，各ノードが対象とするカテゴリに属するか属さないかを推定し，その再現率，適合率，F 値を評価する．ただし，定義したオペレータの有用性を示すため，表 1 に示すように，適用レベル 1 ～ 4 まで段階的にオペレータを増やすこととした．はじめに適用レベル 1 のオペレータだけを用い，次に適用レベル 2 までのオペレータ，適用レベル 3 までのオペレータ，最後に全てのオペレータを適用することで，順次，多くの属性を生成する．

表 2 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．オペレータを増やすに従い，再現率，適合率，F 値が向上している．

表 2: 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．

	再現率	適合率	F 値
レベル 1	0.419	0.555	0.473
レベル 2	0.544	0.629	0.580
レベル 3	0.707	0.745	0.722
レベル 4	0.731	0.757	0.742

4 まとめと今後の課題

本研究では，データマイニングと社会学の間のギャップを埋めるために必要な研究として，社会ネットワーク分析で用いられている指標を体系的に生成する手法を提案した．提案手法では属性生成の過程を 3 つのステップにわけ，各ステップでオペレータを定義し，それらのオペレータの組み合わせにより属性を生成した．またこの手法をアットコスメのデータセットに適用し，リンクに基づくノードの分類への有効性を示した．

今後の課題としては，提案した手法を別のタスクに適用しその有益性を示していきたいと考えている．

参考文献

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

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

[3] 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.

[2]化粧品に関する女性向けのコミュニティサイトであり，ユーザは化粧品の推奨や，気に入ったメンバーのお気に入りメンバーへの登録などができる．