3F6-5

エンティティのネットワーク構造を用いた属性生成

Feature Generation Using Network Structure of the Entity

唐門 準[*1] 松尾 豊[*2] 石塚 満[*1]
Jun Karamon Yutaka Matsuo Mitsuru Ishizuka

[*1]東京大学 情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

[*2]産業技術総合研究所
National Institute of Advanced Industrial Science and Technology

There have been lots of attempts on the aggregation of attributes for advanced relational data mining. On the other hand, an increasing number of studies try to deal with social network data. Sometimes we build some attributes based on the network structure, and in other times we rely on the existing analytical methods such as centrality and clustering from sociology. In this study, we try to bridge the gap between the data mining community and social network analysis through aggregation-based feature generation. The notable feature of our algorithm is the ability to invent several indices that are well studied in social network analysis studies in sociology. We define several general operators that can be applied to a graph structure. Then some combinations of the operators correspond to the traditional indices from the sociology, while other combinations are quite new and sometimes useful. We apply our method to Cora dataset, and show the effectiveness of our approach.

1. はじめに

Web をはじめとする実世界のデータはネットワーク構造を持っている．また近年，データマイニングという言葉が盛んに聞かれるようになっている．これまでのデータマイニングでは，個々のエンティティを独立して扱い，エンティティの各属性に着目することで新たな知識発見などの分析が多く行われてきた．しかし，Web のリンク構造やソーシャルネットワーキングサービス (SNS) におけるユーザのネットワーク，ゲノムや細胞のたんぱく質のネットワーク等のデータは，個々のエンティティが他のエンティティとネットワーク構造を作り出しており，これらのネットワーク構造が分析における重要な鍵となる．

論文を例として考えると，論文はタイトル，発表年など様々な属性値を持っている．また，他の論文との間には引用関係があり，また論文と発表者の間の著者関係なども考えられ，論文や著者との間でネットワーク構造が作りだされている．このネットワークを用いることで，分類やメタデータの付与などのタスクに対する精度を上げることができる．

また一方で，社会ネットワーク分析 [7] と呼ばれる分野では，これらのネットワーク構造を用いて各エンティティを評価するための指標として，中心性，構造同値，クラスタリング係数，パス長などが特に知られている．これらの研究は非常に古くから行われており，近年でも SNS をはじめとする様々なデータを対象に行われているということは，これらのネットワーク構造を分析することの重要性を示していると考えられる．

本稿では，社会ネットワーク分析で使われている指標を分析し，いくつかの段階に分けてネットワーク構造を分析するための様々なオペレータを定義することで，社会ネットワーク分析における様々な指標を自動的に導き出すための手法を提案する．また Cora データベースの論文データを用いた論文ネットワークに対して本手法を適用し，この手法の有益性を示す．

連絡先: 唐門 準，東京大学大学院　情報理工学系研究科　石塚研究室，〒 113-8656 東京都文京区本郷 7-3-1，03-5841-6774，karamon@mi.ci.i.u-tokyo.ac.jp

まず 2 章では本研究と関係する関連研究について紹介する．次に 3 章では，社会ネットワークで用いられている属性を自動的に生成する手法について説明する．4 章では，3 章で提案した手法を実際に論文データに対して適用した結果について述べる．最後に，5 章ではまとめと今後の課題について述べる．

2. 関連研究

2.1 関係学習

関係学習に関する研究の例として，Probabilistic Relational Models(PRM)[3] や Statistical Relational Learning(SRL) らが挙げられる．PRM とは，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するものである．

さらに Alexander らの研究 [6] では，SRL において，関係データから自動的に関係を用いた属性を生成する手法を提案している．また Claudia ら [5] は関係の概念の階層構造を作ることで，関係データに対し，適切に関係性を用いた属性を生成する手法を提案している．

2.2 社会ネットワーク分析に関する研究

本節では，社会ネットワーク分析によって用いられている指標 [8] についていくつかの例をあげて説明する．

ネットワーク密度 ネットワークにおいてノード同士の関係がどれくらい密接であるかを表す指標で，ネットワーク中のリンク数を，最大可能なリンク数で割ったものである．

クラスタリング係数 C で表される指標で，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 x と E_x をあわせたノード集合に対して，ネットワーク密度を考えたものがこの値になる．

次数中心性 ノード数 N のネットワークにおいて，ノード i から隣接するノードへのリンク数を次数といい，k_i で表す．このとき中心性の値が高いほど社会ネットワークにおいてより多くの他者とのかかわりがあることを表す．k_i を最大可能なリンク数 $N-1$ で割った値で求められる．

近接中心性 近接中心性とは，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り，ノード数 N のネットワークにおいてノード i とノード j の間の距離を d_{ij} とすれば，ノード i の近接中心性 C C_i は次の式で表すことができる．

$$C_i^C = \frac{N-1}{\sum_{j \in G} d_{ij}}$$

媒介中心性 媒介中心性とは，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．ノード数 N のネットワークにおいて，ノード j とノード k 間の最短パスの数を n_{ij}，そのうちノード i を通るノード j とノード k の最短パスの数を $n_{ij}(i)$ とすると，ノード i の媒介中心性 C_i^B は次の式で表すことができる．

$$C_i^B = \frac{\sum_{j<k \in G} n_{jk}(i)/n_{jk}}{(N-1)(N-2)}$$

平均パス長 L で表される指標で，ネットワーク中のノード集合からすべての 2 つのノードの組み合わせを作り，それぞれ最短パス長をもとめて平均したものである．

Backstorm らのグループの最近の研究 [2] では，DBLP のデータを用いて研究者をノードとし研究者の共著関係をリンクとした研究者ネットワークを構築し，ネットワークの情報を用いた属性を作り，研究者の発表学会を予測するのに重要な属性を取り出そうとする研究が行われている．その中では学会に所属するメンバー間のリンク関係などコミュニティに関する 11 の属性と，この研究者の共著関係など，研究者自身のネットワーク構造に関する 8 の属性をそれぞれの研究者ごとに生成し，それらを決定木の素性として与えることで，研究者がその学会で発表をするかどうかを推定している．これらの結果では，研究者と直接の知り合い関係にありかつある学会に所属している研究者の数 k が増えるほど，研究者はその学会で発表しやすい傾向が得られ，また知り合い関係のある研究者同士が互いに知り合いであるほどその学会で発表をしやすい傾向が見られている．

2.3 関連研究のまとめ

Backstorm らの研究にみられるように，古くから研究が行われている社会ネットワーク分析の様々な指標は，ネットワーク分析において非常に有効な指標となるといえる．

一方，関係学習の研究分野では，関係データから自動的にオブジェクト間の関係を用いた属性を生成する手法が提案されている．

これらのことからも社会ネットワーク分析で用いられている指標を自動的に生成することは非常に有意義であると考えられる．そこで次章ではこれらの手法について提案する．

3. 提案手法

本章では社会ネットワーク分析で用いられている属性を自動的に生成するための手法について述べる．以下では属性生成の過程を以下の 3 つのフェーズに分割し，各フェーズにおいて必要なオペレータを定義する．

フェーズ 1 対象とするノードを決定するオペレータを定義する．

フェーズ 2 フェーズ 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関係を調べるオペレータを定義する．

フェーズ 3 フェーズ 2 の結果を集計し社会ネットワーク分析において用いられている指標を得るオペレータを定義する．

これらの 3 フェーズのオペレータを組み合わせることで様々なオペレータを得る．

3.1 ノード集合の決定

本節ではノード集合を決定するためのオペレータを定義する．オペレータはネットワーク構造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に分けることができる．

- **ネットワーク構造に基づくノード集合**

ノード x のネットワーク構造を用いた属性を生成するため，ノード集合を決定することを考える．例えば x の隣接ノードをノード集合として考えるような場合である．これは言い換えると，ノード x から 1 ホップのノード集合と考えることもできる．ノード x から 2 ホップ，3 ホップ先のノード集合も同様に考えることで決定することができ，このようなノード集合を次のように表す．

- $N^{(k)}(x)$: k ノード x から k ホップ離れたノード集合

但し $N_x^{(0)}$ はノード x 自身を表す．これを用いることで一般にノード x から k ホップ以内にあるノード集合は次のように表せる．

$$C^{(k)}(x) = N^{(0)}(x) \cap N^{(1)}(x) \cap \ldots \cap N^{(k)}(x) \qquad (1)$$

- **属性値に基づくノード集合**

ノードの属性値によってノード集合を考えることも可能である．例えば，論文ネットワークにおいて，論文があるカテゴリーに所属する論文集合をつくるなどノードのある属性がある決まった値をとるノード集合を考えるものである．このようなノード集合を正のノード集合と呼び，N_p と表す．

この集合 N_p と，前節で述べた集合 $C_x^{(k)}$ 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．

以上 2 種類のノード集合のうち，本稿では社会ネットワーク分析における指標を出すにあたって，以下の 5 つのオペレータを考えることにする．

- $N_x^{(0)}$: ノード x 自身
- $C_x^{(1)}$: ノード x の隣接ノード集合
- $C_x^{(\infty)}$: ノード x から到達可能なノード集合
- $N_p \cap C_x^{(1)}$: 正のノード集合のうちノード x と隣接しているノードの集合
- $N_p \cap C_x^{(\infty)}$: 正のノード集合のうちノード x から到達可能なノードの集合

3.2 リンク関係判別オペレータ

リンク関係判別オペレータとは，3.1 節で述べたオペレータにより決定されたノード集合に対して任意の 2 つのノード集合を取り出し二つのノードの間にある関係が存在するかを調べるものであり，以下のものを定義する．

まず，任意の二つのノード x，y の間に k ホップの関係があるかどうかを調べるオペレータを次のように定義する．

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

また任意の二つのノード x，y の間の距離を求めるオペレータを次のように定義する．

$$t(x,y) = \arg\min_k \{s^{(k)}(x,y) = 1\}$$

同様に任意の二つのノード 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．

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

オペレータとして定義されるものは，上記で挙げたもの以外にも表 1 に挙げたようなものが考えられる．

ここで定義したオペレータは，2 つのノードを対象としているが，実際にノード集合が 2 つ以上の場合，ノード集合からすべてのノードペアをつくり，それぞれについて上記の関係を調べることで，それらの結果をリストとして返す．

3.3 社会ネットワーク分析の指標を求めるオペレータ

3.2 節で定義したオペレータを用いることで得られたリストを以下で定義するようなオペレータで処理することにより，最終的に社会ネットワーク分析で用いられている指標を属性として得ることが可能になる．本稿で定義するオペレータは以下の 4 つである．

- Sum：フェーズ 2 で得られたリストの和
- Ave：フェーズ 2 で得られたリストの平均
- Min：フェーズ 2 で得られたリストの最小値
- Max：フェーズ 2 で得られたリストの最大値

3.4 その他のオペレータ

前節までのオペレータに加え，割合を取るオペレータを定義する．これは 3.1 節で述べた $C_x^{(k)}$ の集合に対して正のノード集合である $N_p \cap C_x^{(k)}$ を考えることができ，それら 2 種類のノード集合に対してそれぞれオペレータを適用し属性値を得て，二つの割合をとるオペレータである．例えば，次の式

$$\frac{Ave \circ s^{(1)} \circ (N_p \cap C_k^{(1)})}{Ave \circ s^{(1)} \circ C_k^{(1)}}$$

では，2 つのノード集合のクラスタリング係数の割合を求めることができる．

以上より本章で定義されるオペレータをまとめると表 1 になる．フェーズ 1~3 まで各フェーズごとに 4 つのオペレータを定義した．これにより 3 つのフェーズを経ることで，$4 \times 4 \times 4 = 64$ 個の属性を得ることができる．

3.5 生成される属性の例

この節では実際に上記で述べたオペレータを用いることで生成することができる社会ネットワーク分析における指標の例を挙げる．

- クラスタリング係数：$Ave \circ s^{(1)} \circ N_x^{(1)}$
- 近接中心性：$Ave \circ t_x \circ C^{(\infty)}(x)$
- 媒介中心性：$Sum \circ u_x \circ C^{(\infty)}(x)$
- 構造空隙：$Ave \circ t \circ N^{(1)}(x)$

4. 実験結果

本章では，論文データを用いて論文の分野を推定する問題を考える．

4.1 実験データ

対象としたデータは，MaCallum らが公開している Cora のデータベースである [4]．コンピュータサイエンスに関する論文約 3 万件を集めたものであり，論文は 69 の研究分野に分類されている．各論文にはタイトル，著者，ジャーナル，引用関係，発表年に関する書誌情報がある．引用関係の数は 5 万件程度，著者は 2 万人程度となっている．このデータから，論文ネットワークを構築した．ネットワークのノードとして，論文エンティティと，著者エンティティを定義し，エッジとして論文エンティティの間の引用関係と，論文エンティティと著者エンティティの間に著者関係を考えた．また各エンティティのもつ属性としては論文エンティティに発表年，研究分野の属性を与えた．

4.2 実験方法

前節で述べたデータを用いて実験を行った．実験は表 1 にある Type1，2，3 にわけて 3 章で定義した各オペレータを適用し，各エンティティに対してネットワーク構造を用いた属性を生成した．ただし実験の対象とするノードは作成した論文ネットワークにおいて，Artificial Intelligence のカテゴリ内の Machine Learning のうち Neural Networks の研究分野に属する論文に対してひとつでもリンクをもっている，あるいはそのカテゴリに属している論文を対象とした．用いた論文ノード集合のうち対象とする研究分野に属する論文 (正例) 数は 781 件，属しない論文 (負例) 数は 901 件であった．論文ネットワークにおける引用関係のリンクは方向を区別せず両方向リンクとして考えた．これらを元に生成された属性を素性として決定木を作り各ノードが，特定の研究分野に属するか属しないかを判別し，Type1~3 になるにしたがって，再現率，適合率，F 値がどのように変化するかを測定した．評価は 10-交叉検定で行った．

決定木の生成に当たっては c4.5[1] を用いた．

本実験において属性の生成にあたり，計算量を減らすため，ノード x から到達可能なノード集合 $C_x^{(\infty)}$ をノード x から 2 ホップ以内のノード集合 $C_x^{(2)}$ に制限した．

4.3 実験結果

上記の条件で実験した結果を表 2 に示す．

結果より，Type1~3 にするに従い F 値が上昇していることがわかる．また Type 3 の際に決定木の上位に来た属性を図.1 にしめす．この結果より，媒介中心性などが上位に来ていることがわかる．

表 1: オペレーター一覧

type	Notation	Input	Output	description
1	$C_x^{(1)}$	node x	a set of nodes	adjacent nodes to x
2	$C_x^{(\infty)}$	node x	a set of nodes	reachable nodes from x
3	$N_p \cap C_x^{(1)}$	node x	a set of nodes	all positive nodes adjacent to x
3	$N_p \cap C_x^{(\infty)}$	node x	a set of nodes	all positive nodes reachable from x
1	$s^{(1)}$	a set of nodes	a list of values	1 if connected, 0 otherwise
1	t	a set of nodes	a list of values	distance between a pair of nodes
2	t_x	node x and a set of nodes	a list of values	distance between node x and other nodes
2	u_x	node x and a set of nodes	a list of values	1 if the shortest path includes node x, 0 otherwise
1	Ave	a list of values	a value	average of values
1	Sum	a list of values	a value	summation of values
1	Min	a list of values	a value	minimum of values
1	Max	a list of values	a value	maximum of values
3	$Ratio_p$	two values	value	ratio of value on positive nodes($N_p \cap C_x^{(k)}$) by the all nodes ($C_x^{(k)}$)

表 2: 実験結果

	再現率	適合率	F 値
type 1	0.4069	0.6256	0.4859
type 2	0.5390	0.5763	0.5429
type 3	0.7640	0.7376	0.7483

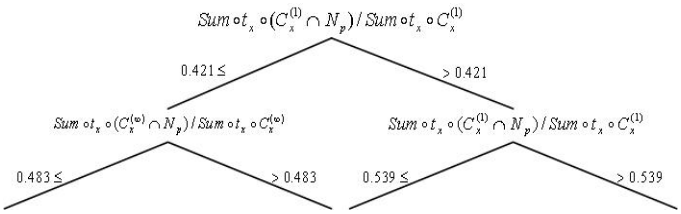

図 1: type 3 の決定木の上位属性

5. まとめと今後の課題

本稿では，社会ネットワークデータから自動的に社会ネットワークで用いられている属性を得る手法を提案した．また実際に Cora データベースにある論文データを用いて，論文ネットワークを生成し，実際に提案した手法を適応して属性を生成し F 値などを評価することで提案手法の可能性を示した．

今後の課題としては，提案した手法を別のデータについて適用することで，さらに手法の有益性を示すことが必要だと考えられる．また社会ネットワーク分析で用いられている指標を分析し，新たなオペレータを組み込み，整理することが考えられる．さらには実験結果を分析して生成された属性のうちどの属性が特に有益であるかを分析していきたい．

参考文献

[1] *http://www2.cs.uregina.ca/ dbd/cs831/notes/ml/dtrees /c4.5/tutorial.html.*

[2] Lars Backstrom, Dan Huttenlocher, and Jon Kleinberg. Group formation in large social networks:membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 44–54, 2006.

[3] Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. Learning probabilistic relational models. *IJCAI*, pp. 1300–1309, 1999.

[4] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval Journal*, Vol. 3, pp. 127–163, 2000. www.research.whizbang.com/data.

[5] Claudia Perlich and Foster Provost. Aggregation-based feature invention and relational concept classes. *In Proceedings of the Ninth SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 167–176. ACM Press, 2003*, 2003.

[6] Alexandrin Popescul and Lyle H. Ungar. Statistical relational learning for link prediction. *IJCAI03 workshop on learning statistical modeling from relational data*, 2003.

[7] 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.

[8] 安田雪．実践ネットワーク分析－関係を解く理論と技法．新曜社, 2001.