5M-3

Mixing Matrix に基づく階層付きカテゴリカルデータの可視化法

伏見　卓恭†　　斉藤　和巳†　　武藤　伸明†　　池田　哲夫†　　大久保　誠也†

†静岡県立大学経営情報イノベーション研究科

1　はじめに

企業間あるいは業種間の取引関係や産業間の投入量など，多くの場面で人や企業などのオブジェクト間にインタラクションが存在する．このようなオブジェクト間あるいはカテゴリ間の関係を効果的に可視化することは，オブジェクト群の全体構造や法則性を把握するのに有用な方法のひとつであると考えられている．一般に関係の強さを表す重み，あるいは，相互関係の量を表す重みなどがあり，オブジェクト間の関係は均一でない場合が多い．このような重み付きのオブジェクト間関係を効果的に可視化する手法として，球面可視化法があげられる [1]．球面可視化法は，2 つのオブジェクト集合を同心円状に配置し，重み付きの隣接関係が類似するオブジェクト同士を原点から見て同一方向に配置する．しかし，この手法では単純に重みの絶対量を用いるため，重みの格差が大きい場合，大きな値の重みが支配的となり，解釈可能な可視化結果を得ることが困難な場合がある．本稿では，上記の問題に対応するように重みを調整し，オブジェクト間の関係を効果的に可視化する方法を提案する．

2　提案手法

提案手法のアルゴリズムとその要素技術について説明する．提案手法はオブジェクト集合 V，オブジェクト間のつながりを表すリンク集合 E が与えられた際，オブジェクトをそのカテゴリに基づき統合し，以下の手順によりカテゴリの布置座標ベクトル群 $\mathbf{X}$ を決定する．

1. リンク集合 E から Mixing Matrix c_{ij} を構築；
2. Mixing Matrix の要素 c_{ij} の Z スコア z_{ij} を計算；
3. Z スコアの双曲線正接関数値 w_{ij} を計算；
4. （3）で計算した重みを用いて，球面可視化法によりカテゴリの布置座標 $\mathbf{X}$ を計算；

(1) から (3) のステップにより，本質構造が規模の格差に隠れてしまうことを回避する．以下の小節で，ステップ (1) および (2) について説明する．ステップ (3) に関しては，重みの格差を是正するためにはシグモイド関数が考えられるが，Z スコアは負の値にも意味があるため，より一般的な双曲線正接関数 (tanh) を用いる．ステップ (4) の球面可視化法については文献 [1] を参照されたい．

2.1　Mixing Matrix

Mixing Matrix は，カテゴリに属するオブジェクト間のリンク数を用いて，どのカテゴリに属するオブジェクト間にリンクが多く存在するかの確率を要素とする行列である [2]．表 1 は，企業をオブジェクト，企業間の取引をリンクとする業種間の Mixing Matrix の例である．

表 1: Mixing Matrix の例

c_{ij}	製造	小売	サービス	卸売	a_i
製造	0.08	0.1075	0.0075	0.055	0.25
小売	0.0704	0.0946	0.0066	0.0484	0.22
サービス	0.16	0.215	0.015	0.11	0.5
卸売	0.0096	0.0129	0.0009	0.0066	0.03
b_i	0.32	0.43	0.03	0.22	1

この行列 $\mathbf{C} = \{c_{ij}\}$ を Mixing Matrix と呼び，カテゴリ数が K の場合，$K \times K$ の行列となる．表 1 の各要素 c_{ij} は，カテゴリ i に属するオブジェクトとカテゴリ j に属するオブジェクト間のリンク数の割合である．この値により，カテゴリ間の癒着や依存度などの関係の強さがわかる．また，行と列それぞれの周辺確率分布を $a_i = \sum_{j=1}^{K} c_{ij}, b_j = \sum_{i=1}^{K} c_{ij}$ とする．以下，カテゴリをノードとする．

2.2　Z スコア

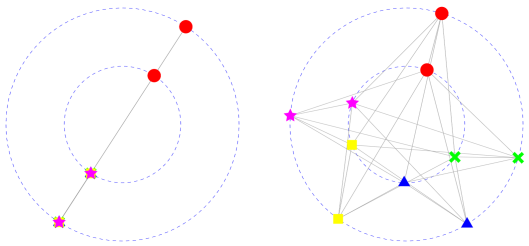
Mixing Matrix の各要素の値 c_{ij} に対して Z スコアを考える．総リンク数 L としたとき，$L \cdot c_{ij}$ の期待値を $L \cdot e_{ij} = L \cdot a_i \cdot b_j$ で計算する．すなわち，オブジェクト間の関係がランダムに決まると仮定したモデルにおける期待値である．$K \times K$ の要素が多項分布に従うと仮定し，ランダムなモデルと比較してどの程度有意に存在するかを表す Z スコアを期待値および標準偏差から計算する．

$$z_{ij} = \frac{L \cdot c_{ij} - L \cdot e_{ij}}{\sqrt{L \cdot e_{ij}(1 - e_{ij})}} \tag{1}$$

Visualization method based on the Mixing Matrix for categorical and hierarchical data

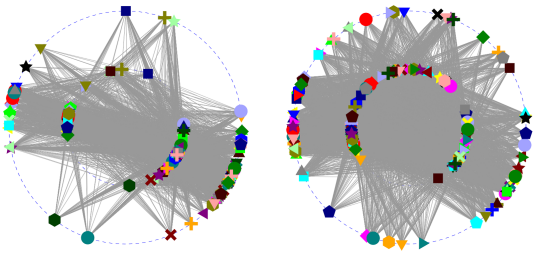
†Takayasu FUSHIMI　†Kazumi SAITO　†Nobuaki MUTOH　†Tetsuo IKEDA　†Seiya OKUBO

†Graduate School of Management and Information of Innovation, University of Shizuoka



(a) 単純法　　(b) 提案法

図 1: 人工データ



(a) 単純法　　(b) 提案法

図 2: 産業連関表データ

Z スコアが正で大きいほど，カテゴリ i のオブジェクトとカテゴリ j のオブジェクト間にリンクが統計的有意に存在するといえる．Z スコアを用いることにより，出現頻度の少ないリンクであっても，特徴的な関係は大きな値となり，規模の格差により隠れてしまう本質構造を考慮できる．

3　評価実験

提案手法に対して，人工データおよび実データを用いて評価する．評価に際して提案法の優位性を確認するために，重みをそのまま使用する単純法と比較する．

3.1　使用データ

1 つ目のデータは，以下のような重みで隣接するノード群からなる人工データである．

表 2: ノード間の重み（人工データ）

階層 1＼2	○	×	△	□	☆
○	10000	1	0.0	0.0	0.0
×	0.0	1	1	0.0	0.0
△	0.0	0.0	1	1	0.0
□	0.0	0.0	0.0	1	1
☆	1	0.0	0.0	0.0	1

2 つ目のデータは，経済産業省のホームページから取得した，2010 年における部門数 80 の産業連関表のデータである．

3.2　実験結果

人工データに対する結果を図 1 に示す．図 1(a) より単純法では，重みの大きな赤い○ノード間が強調され，その他のノード間の関係は全て反対側へ重なってしまうことがわかる．つまり，○ノード以外のノード間関係は重なっている．図 1(b) より提案法では，全体がバラつき，不鮮明だったノード間関係が視覚的に捉えることができることがわかる．重みの絶対量が大きな赤い○ノード間は，原点からみて同一方向に（なす角度が比較的小さい位置に）配置されており適切に可視化されていることがわかる．他のノードに関しても，重みの絶対量が少ないながらも隣接しているノード同士は原点から見て比較的同一方向に配置されている．人工データを用いた評価実験により，提案法は重みの絶対量が大きなペア関係を維持しながら，規模の格差により隠れてしまう本質構造を可視化できることが示唆された．

産業連関表に対する結果を図 2 に示す．図 2(a) の単純法では，重みの大きな同部門間が強調され，左側に小売・サービス系の部門，右側に製造・工業系の部門が集中し，直線状に重なって配置されている．つまりノード間関係は不鮮明である．図 2(b) の提案法では，全体がバラつき，不鮮明だったノード間関係が視覚的に捉えることができる．バラつきがある分，左上に公共系，左下に情報系のサービス部門や上側に化学系，右側に電子系，右下に通信系の工業部門など詳細な部門間関係が把握可能になった．

4　おわりに

本稿ではオブジェクト間あるいはカテゴリ間の相互関係を効果的に可視化する方法を提案した．人工データおよび産業連関表を用いた実験結果から，提案法は規模の大きさにより隠れてしまう関係の本質構造を可視化できることがわかった．今後は，より多様なデータを用いて評価し，提案法の有効性・有用性を検証していきたい．

謝辞　本研究は，株式会社豊田中央研究所との共同研究および，科学研究費補助金基盤研究 (C)(No.23500128) の補助を受けた．

参考文献

[1] 久保田大和，伏見卓恭，斉藤和巳，風間一洋：重み付き 2 部グラフの可視化法，ネットワークが創発する知能研究会 (Jwein'11) (2011).

[2] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, Vol. 67, No. 2, pp. 026126+, Feb 2003.