F-053

二部グラフにおけるクラスタリングアルゴリズムの比較

Comparing Clustering Algorithms in Bipartite Networks

原田　恵雨†　鈴木　育男†　山本　雅人†　古川　正志†

Keiu HARADA　Ikuo SUZUKI　Masahito YAMAMOTO　Masashi FURUKAWA

1　はじめに

近年の WWW の発達に伴って，情報をクラスタとして抽出する需要が高まっている．これは膨大な情報を扱うために，高速で精度の高いクラスタを抽出するためのアルゴリズム開発が要求されている．こうした情報の中にはネットワークを形成するものが存在している．ネットワーク中のリンクが密なノード集合を抽出する手法は，Newman ら [1] の *Modularity* を用いた手法が有名であり，*NewmanFast* と呼ばれている．ここで，抽出されたノード集合はコミュニティと呼ばれている．*NewmanFast* は高速であるが，*Modularity* の高いコミュニティは抽出できない欠点を備えている．また，*Modularity* を用いた手法は，二部グラフのようなリンクの張り方に制約のあるネットワークを扱うことができない．クラスタリングの対象となるネットワークには二部グラフの構造を持つものが少なからず存在している．二部グラフ構造に対応したクラスタリング手法として著者ら [2] は，順序最適化による手法を提案している．

本研究では，二部グラフを対象としたクラスタリングアルゴリズムの比較を目的とし，順序最適化によるクラスタリングアルゴリズムの精度を調査する．

2　関連研究

さまざまなクラスタリングアルゴリズムの比較研究として，Danon ら [3] の研究がある。彼らはあらかじめクラスタ構造が分かっているネットワークに対して，ランダムリンクを変化させたときの元のクラスタ構造に対する正規相互情報量 (*normalized mutual information, NMI*) を計測し，比較を行っている．以下に *NMI* の計算式を示す．

$$I(A,B) = \frac{-2\sum_{i=1}^{N_M^A}\sum_{i=1}^{N_M^B} n_{ij}^{AB}\log\left(\frac{n_{ij}^{AB}S}{n_i^A n_j^B}\right)}{\sum_{i=1}^{N_M^A} n_i^A \log\left(\frac{n_i^A}{S}\right) + \sum_{j=1}^{N_M^B} n_j^B \log\left(\frac{n_j^B}{S}\right)} \quad (1)$$

ここで A,B はそれぞれ，あらかじめ分かっているクラスタ集合，得られたクラスタ集合を表す．n_i^A は A の i 番目のクラスタ内のノード数，n_j^B は B の j 番目のクラスタ内のノード数，n_{ij}^{AB} は A と B で共通するノード数，S は全ノード数である．

得られたクラスタリング集合のモジュール度を測る評価値としては，Newman ら [1] が提唱する *Modularity* がある．*Modularity* は与えられたネットワークに対してクラスタに分けたとき，同次数を持つランダムなクラスタを仮定した場合の内部リンク率を引いた値となる．*Modularity* は以下の計算式で示される．

$$Q = \sum_{s=1}^{S}\left(\frac{l_s}{M} - \left(\frac{k_s}{2M}\right)^2\right) \quad (2)$$

ここで，S はクラスタ数，l_s は s 番目のクラスタの内部リンク数，k_s は s 番目のクラスタの総次数，M はネットワークの総リンク数である．

NMI を用いて評価を行う場合，あらかじめ正解のクラスタを定めておく必要がある．しかし，複数のクラスタにまたがるリンクが多数存在する場合，正解のクラスタを定められない短所を持つ．それに対して，*Modularity* はクラスタ数に依存しない長所がある．しかし，[4] のような問題がある．

3　クラスタリングアルゴリズム

二部グラフに対するクラスタリングアルゴリズムは，一部グラフのそれと比較して少ない．Co-clustering[5] は隣接行列の第二固有ベクトルを用いて変換し，k-means 法でクラスタリングする手法である．また，WeakestPair[6] は正規化隣接行列を用いた類似度を定義し，もっとも類似度の低いノード間リンクを切除していく手法である．また，[7] では，二部グラフ用に拡張した *Modularity* を用いてクラスタリングを提案している．

著者ら [2] は順序最適化によるクラスタリングアルゴリズムを開発した．これは，二部グラフの各ノード間にコストを定義し，コストを用いた TSP を解くことで計算量を抑えた手法である．得られた TSP 解より，クラスタ間の境界となるコストを決定することでクラスタを抽出する．

4　実験

4.1　明示的なクラスタ構造を持つ二部グラフモデル

クラスタリングアルゴリズムの質を定量的に評価し，比較するために，明示的なクラスタ構造を持つランダムネットワークを実験材料として採用した．

[7] では明示的なクラスタ構造を持つ二部グラフモデルとして，アクターと所属するチームからなる二部グラフをランダムに生成する方法を提案している．生成手順を以下に示す．

1. アクターを各モジュール k に対し S_k 人によって構成される，N_M 個のモジュールに振り分け，モジュールの“色”を決める．
2. チーム a を生成し，チームの構成人数 m_a とチームの“色”を決める．“色”はアクターのモジュールに決められた“色”の中より選択する．
3. 各チームに対して以下の手順を行う．
   (a) 同色選択確率 p によって同じ色のモジュールに属するアクターにリンクを張る．
   (b) 異色選択確率 $1-p$ によって異なる色のモジュールに属するアクターにリンクを張る．

† 北海道大学大学院　情報科学研究科　複合情報学専攻

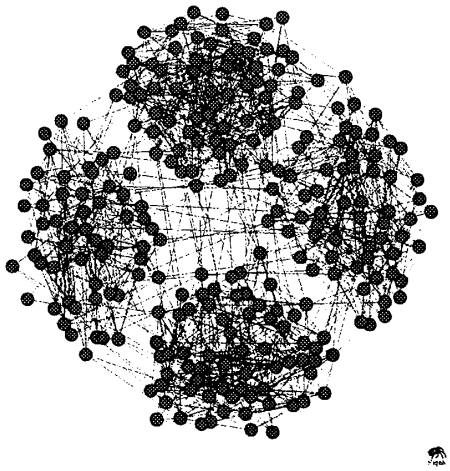

図 1: *TAM*(4, 32, 128, 14, 0.95, 1)

4. 手順 3 にしたがってチームを N_T 個生成する.

ここで, S_k の決め方を $h = S_{i+1}/S_i$ とすることで, モジュールサイズの非対称性を得ることができる.

このアルゴリズムによって生成されたグラフを以下では $TAM(N_M,S_S,m,N_T,p,h)$ と記す.

図 1 に *TAM*(4, 32, 128, 14, 0.95, 1) を示す.

4.2 クラスタ構造をもつネットワークに対する比較実験

実験の手順を以下に示す. これは [7] の実験に対して順序最適化によるクラスタリングアルゴリズムを適用し *NMI* を計測したものである.

1. *TAM* のパラメータ N_M, S_S, m, N_T を固定する.
2. p を変化させ, 順序最適化によるクラスタリングアルゴリズムを適用する.
3. 得られたクラスタ集合に対して *NMI* を計測する.

順序最適化をするときのノード間コスト $Cost(x,y)$ を以下のように定めた.

$$Cost(x,y) = -\frac{|A_x \cap A_y|}{\sqrt{|A_x||A_y|}} \tag{3}$$

ここで A_x, A_y はそれぞれ x, y の隣接ノード集合とする. また, クラスタの境界となるコストを以下のように定めた.

$$CutCost = (cost_{max} - cost_{min}) \times r + cost_{min} \tag{4}$$

ここで $cost_{max}$, $cost_{min}$ はそれぞれ得られた経路に対するノード間コストの最大値, 最小値である. また, r は $0 \le r \le 1$ の実数とする.

5 結果と考察

表 1 に計測した *NMI* の値を示す.

この結果より, [7] で示されている値よりも低いことが分かる. これは, 式 4.2 により選択した境界コストにより, 本来 1 つであるクラスタが複数に分割されてしまっていることや, 複数のクラスタが 1 つになっていることが調査の結果判明した. また, p によって *NMI* が最大となる境界コストが異なることも分かった. したがって, 境界コストの選択は慎重に行わなければならないことが分かった.

表 1: 計測した *NMI* の値

p	*actor*, r=0.8	*team*, r=0.8	*actor*, r=0.9	*team*, r=0.9
0.20	0.0391	0.0675	0.0360	0.00580
0.25	0.126	0.0967	0.0602	0.0334
0.30	0.157	0.144	0.0423	0.0212
0.35	0.172	0.199	0.0713	0.0858
0.40	0.0533	0.0235	0.000	0.000
0.45	0.101	0.161	0.0531	0.0568
0.50	0.110	0.101	0.0126	0.0188
0.55	0.0780	0.0783	0.0253	0.0395
0.60	0.191	0.125	0.131	0.0632
0.65	0.231	0.0524	0.116	0.0124
0.70	0.160	0.0868	0.0845	0.0213
0.75	0.358	0.174	0.174	0.0930
0.80	0.425	0.0921	0.384	0.0148
0.85	0.275	0.0771	0.146	0.0218
0.90	0.343	0.0675	0.240	0.0253
0.95	0.476	0.0866	0.305	0.0454

6 まとめ

本研究では, 順序最適化によるクラスタリングアルゴリズムの精度を *NMI* によって定量的に評価した. 境界コストの選び方によってクラスタが多数に分割されることにより, *NMI* の値が小さくなってしまう事を確認した. この結果より, 境界コストの選び方を決定する方法を発見する課題が与えられた.

参考文献

[1] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, Vol. 69, p. 026113, 2004.

[2] 原田恵雨, 吉井伸一郎, 古川正志. 二部グラフ分割問題の順序最適化としての解法. ロボティクス・メカトロニクス講演会 Robomec '07 予稿集, pp. 2A1–J08, 2007.

[3] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 9, pp. 8–+, sep 2005.

[4] Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *PNAS*, Vol. 104, No. 1, pp. 36–41, January 2007.

[5] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Knowledge Discovery and Data Mining*, pp. 269–274, 2001.

[6] 石田和成. 潜在的ウェブログコミュニティ抽出のための二部グラフ分割アルゴリズム. 人工知能学会 SIG-SWOA404-01, 2004.

[7] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral. Module identification in bipartite and directed networks. *arXiv:physics/0701151*, Jan 2007.