D-004

ハイパーグラフクラスタリングにおけるエッジ追加順序の比較
Comparison of Edge Addition Orders in Hypergraph Clustering

伊藤 柊太 † 伏見 卓恭 †
Shuta Ito Takayasu Fushimi

1. はじめに

グラフ理論において一般的に扱われるグラフは，2 ノード間の隣接関係を表すことができる．ハイパーグラフはグラフの概念を一般化したもので，任意の数のノード間の隣接関係を表すことができるため，その表現能力から近年注目を浴び多くの研究がなされている．代表例として，ハイパーグラフにおけるコミュニティ抽出，あるいは，ノードクラスタリングが挙げられる．通常のグラフのクラスタリング手法を一般化することで，ハイパーグラフに適用する手法が多く提案されているが，精度，計算量の観点から代表的な手法は未だ確立されていない．

本研究では，ハイパーグラフにおけるノードとハイパーエッジの接続関係を二部グラフに展開し，多数のノードとハイパーエッジに跨るようなノイジーな二部グラフエッジを検出する手法を提案する．具体的には，自然言語処理の分野で用いられる TF-IDF に基づき二部グラフエッジに重みを付与し，重みの大きな重要エッジから順に復元するアルゴリズムを提案する．裏を返せば，重みの小さいノイジーなエッジの復元は後回しにされるため，精度良くクラスタを検出できる．そして，クラスタの抽出精度の観点から，関連する重み付け手法との違いを比較する．

2. 関連研究

本研究では，対象のハイパーグラフを二部グラフに展開し，二部グラフエッジを 1 本ずつ追加する凝集型階層クラスタリングの要領でハイパーグラフのノードをクラスタリングする手法を提案する．この節では，ハイパーグラフをクラスタリングする手法を俯瞰し，本研究の位置づけを示す．

グラフに対するスペクトラルクラスタリングやスペクトラル疎性化を一般化し，ハイパーグラフに適用できるようにした手法として文献 [1, 2] がある．これらは，大規模な行列に対して固有ベクトルを求める必要があり，計算時間がかかるという問題がある．グラフに対するモジュラリティ指標を一般化し，ハイパーグラフに適用できるようにした手法として文献 [3, 4] がある．Kumar らは，対象のハイパーグラフを clique 展開 [5] し，clique 展開に伴い増加した各ノードの次数を是正した隣接行列と configuration モデルによる期待エッジ数行列を用いてモジュラリティを計算している．モジュラリティ最大化を達成するクラスタリングを実現するために，Louvain 法 [6] に基づき高速にノードをクラスタリングするフェーズとクラスタリング結果からハイパーエッジのカットを誘導する重みを更新するフェーズを繰り返す．クラスタ間をバランス良く跨ぐようなハイパーエッジの重みを小さくし，次のフェーズでもカットされやすいように更新する．重み行列の更新量が十分小さくなったところで反復を終了し，クラスタリング結果を出力する．Kaminski らは，Chung-Lu モデルをハイパーグラフに適用できるように一般化し，それに基づき厳密なモジュラリティ指標を定義している．モジュラリティ最大化を達成するクラスタリングを実現するために，CNM 法 [7] と同様のアルゴリズムを適用しているが，非常に計算時間がかかるため，実際にはランダムにハイパーエッジを選び追加する試行を複数回行うアルゴリズムで実験している．

3. 提案手法

N 個のノードの集合 $\mathcal{V}$ と M 個のハイパーエッジの集合 $\mathcal{R} \subset \mathcal{P}(\mathcal{V}) \setminus \{\varnothing\}$ からなるハイパーグラフ $H = (\mathcal{V}, \mathcal{R})$ に対して，star 展開 [5] により二部グラフ $B = (\mathcal{V}, \mathcal{R}, \mathcal{E})$ を得る．この二部グラフにおいて，$\mathcal{R}$ はノード集合として扱われ，$\mathcal{E} \subset \mathcal{V} \times \mathcal{R}$ は $\mathcal{V}$ と $\mathcal{R}$ の要素をつなぐ二部グラフエッジの集合 $\mathcal{E} = \{(v, r); v \in r, r \in \mathcal{R}\}$ である．各エッジ $e \in \mathcal{E}$ に対して，TF-IDF に基づき重み $w(e)$ を定義し，この重みが高い順に $L = |\mathcal{E}|$ 本のエッジを順に付与する．ここで，二部グラフエッジ $e = (v, r)$ の重みは以下のように計算される：

$$\begin{aligned} w((v,r)) &= TF(v;r) \cdot IDF(v) \\ &= \frac{f_{v,r}}{\sum_{v' \in \mathcal{V}} f_{v',r}} \cdot \left(\log \frac{N}{g_v} + 1 \right) \end{aligned}$$

ここで，$f_{v,r}$ は二部グラフエッジ $e = (v, r)$ の多重度を表す．元のハイパーグラフで言うと，ハイパーエッジ r にノード v が出現する回数を表している．単純ハイパーグラフ（多重度が 1）の場合は，全ての $(v, r) \in \mathcal{E}$ に対して $f_{v,r} = 1$ となる．また，$g_v = |\{r \in \mathcal{R}; v \in r\}|$ は，ノード v の次数であり，TF-IDF の文脈では単語 v の DF に相当する．

エッジを付与して得られた二部グラフにおいて，可到達であるノード集合が同じクラスタに属していると考えクラスタリング結果を得る．

提案手法では，ノード集合に対してハードクラスタリングを行うため，Disjoint Set を用

†東京工科大学コンピュータサイエンス学部

表 1: ハイパーグラフの基本統計量

データ	$N = \|\mathcal{V}\|$	$M = \|\mathcal{R}\|$	$L = \|\mathcal{E}\|$
Syn1	500	50	2,669
Syn2	500	50	2,815
YouTube	45,352	13,251	120,877

いて B の可到達性を保持することができる．Disjoint Set に対して Union-find アルゴリズムを用いることで，$O(\alpha(L))$ の計算量で可到達でないノードを可到達に変更する操作を行うことができる．このとき，$\alpha(N)$ はアッカーマン関数 $A(N, N)$ の逆関数とする．上述した操作をすべての $\mathcal{E}$ に対して行う場合に計算量が $O(L \cdot \alpha(L))$ となり，これが最悪計算量となる．一般に，$O(\alpha(N))$ は $O(\log N)$ より小さい計算量であることから，$\mathcal{E}$ を重み順に降順ソートする操作がボトルネックとなり，本論文の提案手法によるクラスタリングの計算量は $O(L \log L)$ となる．

4. 評価実験

4.1. データセット

本研究では，以下に示す設定で人工的にハイパーグラフを構築する．人工的なハイパーグラフを作成する際に，クラスタ数 K，1 つのクラスタ $\mathcal{V}_k$ に含まれるノード数 n_k，1 つのクラスタに完全に含まれるハイパーエッジ数 m_k，複数クラスタのハイパーエッジに含まれるノード数 n' の 4 つのパラメータを与える．k 番目のクラスタ $\mathcal{V}_k$ では，ノード番号が $[(k-1)n_k+1, kn_k)$ のノードに対して確率 1/2 で m_k 本のハイパーエッジを選択し，そのハイパーエッジに含まれるようにする．次に，すべてのノードからランダムに n' 個選択し，ノイズノードとする．ノイズノードは，自身が属するクラスタ外のハイパーエッジに確率 1/2 で含まれるようにする．

実データとして，動画共有 Web サイト”YouTube” [‡]のコミュニティ情報を用いる．各コミュニティに所属しているユーザをノード，各コミュニティをハイパーエッジとしてハイパーグラフを構築し，最大連結成分を抽出したものを対象とする．”Stanford Network Analysis Project”(SNAP) [§]のデータセットを利用した．これらのハイパーグラフデータに関する基本統計量を表 1 に示す．

4.2. 比較に用いる手法

提案手法は，star 展開した二部グラフエッジに対して TF-IDF に基づき重みを計算し，重みが大きい順にエッジを復元するアルゴリズムである．TFIDF による重み付け法の有効性を検証するために，以下の重み付け法と比較する．

TF ：TF-IDF の計算に用いる TF 値 $TF(v; r)$ を二部グラフエッジ (v, r) の重みとする．本論文では，多重度のない単純ハイパーグラフを対象とするため，あるハイパーエッジ r に含まれるノード $v \in r$ は全て同一の重み $TF(v; r) = TF(r)$ となる．

IDF ：TF-IDF の計算に用いる IDF 値 $IDF(v)$ を二部グラフエッジ (v, r) の重みとする．

Okapi ：Okapi BM25 における以下に示すスコアを二部グラフエッジ (v, r) の重みとする．

$$\frac{f_{v,r}(k_1+1)}{f_{v,r}+k_1\left(1-b+b\frac{|r|}{\bar{h}}\right)} \cdot \left(\log \frac{N-g_v+0.5}{g_v+0.5}\right)$$

ここで，$\bar{h}$ はハイパーエッジ次数の平均値である．

Random ：$[0, 1]$ の一様乱数を二部グラフエッジ (v, r) の重みとする．

4.3. 重み付け手法の類似度

図 1 は，人工グラフ 1 に対して各指標によって重み付けを行い，その重みによってグラデーションを施した接続行列である．接続行列は横軸をハイパーエッジ，縦軸をノードとしてハイパーグラフを可視化したもので，ハイパーグラフのノード i がハイパーエッジ j に含まれている場合は接続行列の i 行目の j 列目に印が付けられている．今回の実験では，この印に重みが大きいほど薄い色，小さいほど濃い色を着色している．TF-IDF と Okapi BM25 は，クラスタ内に完全に収まっている通常のノードと比較して，1 つのクラスタ内以外のハイパーエッジにも含まれるノイズのような頂点の重みを小さくしている．このような重み付けをすることによって，クラスタ内に収まっているノイズではない頂点から復元されるので，これら 2 つの指標はクラスタリングを比較的正しく行うことができると考えられる．TF は，単純ハイパーグラフにおいては分子の値が 1 固定であることから分母の値のみで決定する．分母はハイパーエッジに含まれるノード数となっているので，同一のハイパーエッジに含まれる頂点はすべて同じ重みとなる．よって，同一の列の印はすべて同じ色となっている．IDF はノードによって決定する値で，多くのハイパーエッジに跨るノードほど小さな重みとなることから，ノイズとなる頂点に小さな重みが割り当てられる．また，ノードによって決定する値なので，接続行列の同一の行の印はすべて同じ色となっている．

人工グラフ 2 の各指標の接続行列も，人工グラフ 1 の接続行列と同じ特徴しか現れなかったため割愛する．

図 2(a)，(b)，(c) はそれぞれ，人工グラフ 1，2 と YouTube データセットに対して TF-IDF を

[‡]https://youtube.com
[§]http://snap.stanford.edu

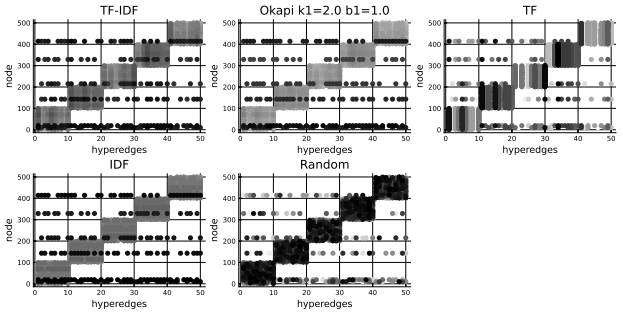


図 1: 人工データ 1 の二部グラフエッジの復元順序グラデーション

用いて二部グラフエッジを重み付けした際の復元順序と，その他の様々な指標を用いた際の復元順序の f1 measure の推移である．横軸は何本のエッジを復元したか，縦軸はすでに復元された二部グラフエッジを 1，そうでない二部グラフエッジを 0 とした L 次元の 2 ベクトルの f1 measure の推移である．図 1 からも予想できる通り，図 2(a)，(b) どちらも，Okapi BM25 との f1 measure が最も高く，またランダムな重み付けとの f1 measure が最も低いことがわかる．図 2(c) では，人工データと比べ，IDF との f1 measure が速い段階では大きくならないことが特徴的である．これは，IDF が YouTube データセットから構築されたハイパーグラフの特徴を捉えられず，すべての二部グラフエッジにおいて IDF の値がほとんど変わらなかったからである．つまり，YouTube のユーザー内で，大量のコミュニティに属しているような人は少ないことを表していると考えられる．

4.4. クラスタリング精度の比較

図 3(a)，(c) はそれぞれ，人工グラフ 1，2 に対して様々なパラメータを設定した Okapi BM25 を用いてクラスタリングを行った際の f1 measure の推移である．少ないエッジの復元で f1 measure が高くなるということは，ハイパーグラフにとって重要な二部グラフエッジに大きな重みを正しく付与しているということであることから，そのような重み付けを行うパラメータが最も優れている．少ないエッジ復元で f1 measure が高くなるパラメータは，図 3(a)，(c) ともに k1=2.0，b=1.0 のときである．以降の比較実験の際はこのパラメータを使うこととする．

図 3(b)，(d) はそれぞれ，人工データ 1，2 に対して実際にクラスタリングを行い，それぞれの結果が人工データの理想のクラスタリング結果との f1 measure を表している．横軸は追加した二部グラフエッジ数，縦軸は f1 measure を表している．指標によって最終的に追加した二部グラフエッジの総数が異なっているが，復元する順序によって，クラスタ数が 1 になるまで復元する際の最小な二部グラフエッジ数が異なるからである．例えば Random によって復元した際は，二部グラフエッジはその重要度にかかわらずまんべんなく復元されるので，他の指標と比べ相対的に早くクラスタ数が 1 になる．図 3(a)，(c) と同様の理由から，少ない復元エッジ数で f1 measure が高くなる指標が優れている．すなわち，グラフ 1，2 ともに本実験で最も優れている指標は TF-IDF である．

実データには正解データが存在しないため，YouTube データセットのクラスタリングの評価は，グラフ分割の質を定量化した Modularity を用いて評価する．図 4 は，YouTube データセットにおける Modularity の推移を表している．横軸が復元した二部グラフエッジの割合，縦軸が Modularity を表している．4.3 節で述べたとおり，TF-IDF と Okapi BM25 は重み付けされた二部グラフエッジの順序の相関が非常に高いため，Modularity も非常に似通った遷移をする．また，TF による重み付けによるクラスタリングの際も，Modularity は非常に似通った遷移をする．これは 4.3 節でも述べたとおり，IDF がすべての二部グラフエッジ内で大きな差がなかったため，TF-IDF と Okapi BM25 ともに，TF の値の影響を大きく受けているためだと考えられる．

5. おわりに

本研究では，いくつかのハイパーグラフを対象に，二部グラフに展開し，エッジに様々な指標を用いて重みを定義，そして重みの降順にエッジを復元することで，クラスタリングに最適な重み付けを検証した．多くのハイパーグラフにおいて，本研究で用いた指標の中では TF-IDF による重み付けが最適であることが分かった．

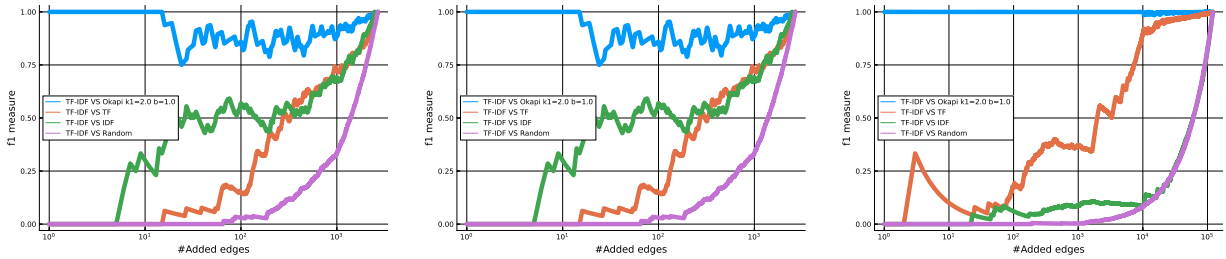
しかし，ノイズノードのクラスタ内の二部グラフエッジの重みがクラスタ外の二部グラフエッジとほとんど変わらなかったため，より良い重み付け手法の考案が今後の課題となる．また，より多様な実データを用いて構築されたハイパーグラフに本研究のクラスタリングを適用し，適切なクラスタを得られるか検証することも今後の課題となる．

謝辞 本研究は，JSPS 科研費（No.20K11940）の助成を受けたものである．

参考文献

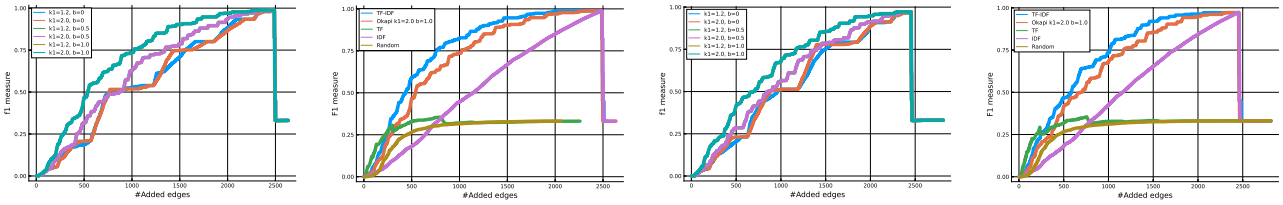
[1] Ahn, K., Lee, K. and Suh, C.: Hypergraph Spectral Clustering in the Weighted Stochastic Block Model, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 12, pp. 959–974 (2018).

[2] Soma, T. and Yoshida, Y.: Spectral Sparsification of Hypergraphs, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, USA, Society for Industrial and Applied Mathematics, pp. 2570—-2581 (2019).



(a) 人工データ 1　(b) 人工データ 2　(c) YouTube データ

図 2: 各重み付け法による二部グラフランキングの類似性



(a) 人工データ 1 に対する Okapi パラメータごと　(b) 人工データ 1 に対する重み付け法ごと　(c) 人工データ 2 に対する Okapi パラメータごと　(d) 人工データ 2 に対する重み付け法ごと

図 3: f1 measure の推移

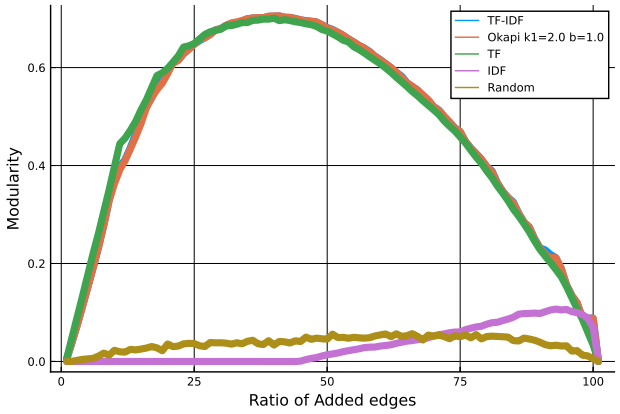


図 4: モジュラリティの推移（YouTube データ）

[3] Kumar, T., Vaidyanathan, S., Ananthapadmanabhan, H., Parthasarathy, S. and Ravindran, B.: Hypergraph Clustering: A Modularity Maximization Approach, *CoRR*, Vol. abs/1812.10869 (2018).

[4] Kamiński, B., Poulin, V., Pralat, P., Szufel, P. and Théberge, F.: Clustering via hypergraph modularity, *PLOS ONE*, Vol. 14, No. 11, pp. 1–15 (2019).

[5] Zien, J. Y., Schlag, M. D. F. and Chan, P. K.: Multilevel spectral hypergraph partitioning with arbitrary vertex sizes, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 18, No. 9, pp. 1389–1399 (1999).

[6] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E.: Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2008, No. 10, p. P10008 (2008).

[7] Clauset, A., Newman, M. E. J. and Moore, C.: Finding community structure in very large networks, *Physical Review E*, Vol. 70, No. 6, pp. 066111+ (2004).