LO-003

グラフ構造に基づくコミュニティ抽出手法
Community Mining Algorithm by Using Graph Structure

高橋 篤[†] 荒井 幸代[‡] 須貝 康雄[‡]
Atsushi Takahashi Sachiyo Arai Yasuo Sugai

1. はじめに

ある共通の話題を種にボトムアップに生成された Web ページの集合を切り出す方法は Web コミュニティマイニングとして近年重要な研究課題である．その方法は，Web の内容による分類とグラフ構造による分類の 2 つに大別できる．

後者の手法は，Web に限らず，物理的な道路網や人間のネットワークにおける重要な部分グラフを抽出する方法としても利用可能である．本研究ではグラフ構造に基づくコミュニティ抽出法として G.W.Flake らが提案した最大フローアルゴリズムを利用した手法 [1][2] を取り上げ，いくつかの実験を通じてこの問題点を指摘し，これを改善するアルゴリズムを提案する．

以下，2 章では本研究におけるコミュニティを定義し，コミュニティ抽出性能の評価規範に考え方を示す．3 章では，グラフ解析に基づく抽出アルゴリズムにおいて広く用いられている最大フローを用いた方法について簡単に説明する．4 章では，Flake の手法の問題点を指摘し，この問題を改善するアルゴリズムを提案する．5 章においてその性能を評価するための実験方法と結果を示し，これを考察する．最後に結論と今後の課題を述べる．

2. 対象問題

Web において各ページをノード，ページ間に張られているハイパーリンクを辺に対応させることにより，グラフとしてとらえることができる．本研究では，「コミュニティとみなす部分グラフ内のノード間の繋がりは外のノードとの繋がりよりも強い.」とする G.W.Flake[1] のコミュニティの定義に従う．

図 1 にこの定義に従った典型的なコミュニティの例を示す．ここでは点線で囲まれたノード集合をコミュニティとみなし，点線の外のノードとの接続辺よりも，コミュニティ内での接続辺を多く持っている．理想的なコミュニティ抽出アルゴリズムは，図 1 の中央の太線のように，コミュニティとコミュニティ以外のノード集合を分けることができなければならない．

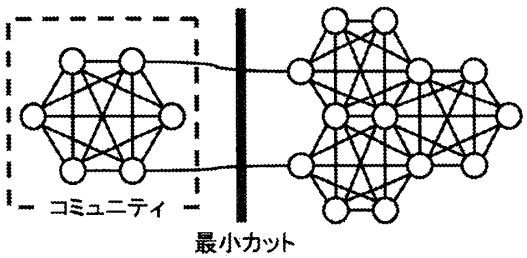

図 1: Web コミュニティの例

3. 最大フローアルゴリズムによるコミュニティ抽出法

最大フローアルゴリズムはノード集合を V，辺集合を E とするグラフ $G=(V,E)$ に対して，ノード u，v 間の有向辺 $(u,v)\in E$ に辺容量 $c(u,v)$，および，2 つのノード $s,t\in V$ を与えたとき，すべての辺で容量を超えることなくソースノード s からシンクノード t へのフローの最大量を求める手法である．ここで s，t をそれぞれソースノード，シンクノードと呼ぶ．最大フローアルゴリズムによって，ソースノードから不飽和辺をたどるとグラフ上の s と t を切り離すことのできる切断辺の最小集合が得られ，この集合を最小カットセットと呼ぶ．カットセットとはソースノード s とシンクノード t を切り離すことのできる辺の集合であり，カットセットに含まれる辺の容量の総和が最小となるものが最小カットセットである．

最大フローアルゴリズムによってグラフの辺を切断し，s から到達可能なノード集合をコミュニティとして抽出するのが従来の一般的方法である．図 2 に最大フローアルゴリズムを用いた既存手法 [1][2] の手順を示す．

入力 初期シードノード集合 $S=\{v_1^s,v_2^s,\ldots,v_n^s\}$
出力 コミュニティ $Com=\{v_1^s,\ldots,v_i^s\}(i=1,\ldots,n)$

1. $k=\mid S\mid$（$\mid\cdot\mid$ は集合・の要素数を表す）.
2. $G=(V,E)$；各 $v_i^s\in S$ から深さ 2 以内の周辺グラフ.
3. すべての $(u,v)\in E$ において辺容量 $c(u,v)=k$ とする.
   $(u,v)\in E$，かつ，$(v,u)\notin E$ のとき，辺容量 $c(v,u)=k$ の辺 (v,u) を E に加える.
4. 仮想ソースノード s，仮想シンクノード t を V に加える.
5. s から各 $v_i^s\in S(i=1,2,\ldots,n)$ への辺 (s,v_i^s) を E に加える． 辺容量 $c(s,v_i^s)=\infty$ とする.
6. すべての $v\in V$（但し，$v\notin S\cup\{s,t\}$）について辺容量 $c(v,t)=1$ の辺 (v,t) を E に加える.
7. s-t 最大フローアルゴリズムを実行する.
8. シードノードから不飽和辺をたどって到達可能なノード集合をコミュニティ Com とする.
9. $n\leftarrow n+1$ とする.
   Com 内の各ノードを次数に基づいてランク付け，このうち高いランクのノードを v_{i+1}^s とし S に加える.
10. 手順 2〜9 を望みの規模のコミュニティになるまで繰り返す.

図 2: 最大フローアルゴリズムを用いた抽出手法 [1][2]

手順 1 で辺容量 k を設定し，手順 2 で周辺グラフの探索をする．シードノードとリンク 1 のノードを深さ 1 のノード，リンク 2 のノードを深さ 2 のノードと呼ぶ．周辺グラフとは，シードノードと深さ 2 までで接続してい

[†]千葉大学大学院工学研究科 建築・都市科学専攻 博士前期課程
[‡]千葉大学大学院工学研究科

る部分グラフである．手順 3 で，ノード間に双方向の辺を張っている．手順 4〜6 では，仮想ソースノードと仮想シンクノードを追加し，仮想シンクノードとシードノード間，仮想シンクノードと周辺グラフのノード間に辺を張る．また，仮想ソースノードに接続しているすべての辺は仮想ソースノードを始点とし，仮想シンクノードと接続しているすべての辺は仮想シンクノードを終点とする．この際，仮想シンクノードとシードノード間の辺容量は ∞ で，仮想シンクノードと周辺グラフのノード間の辺容量は 1 である．手順 7 から最大フローが求まり，手順 8 から最小カットが得られる．最小カットによって切断されるノード集合がコミュニティとして抽出される．

図 3 に，手順 1〜8 の適用後に得られるコミュニティを示す．図中では省略しているが，シードノード以外の各ノードから仮想シンクノードへそれぞれ辺を持つ．また，有向辺の向きも省略しているが，仮想ソースノードと仮想シンクノード以外のすべてのノード間は双方向の辺を持つ．図 3 の点線は切断辺，黒点がコミュニティ内のノードを示している．最小カットに含まれる辺が切断されたあとに残ったノード集合がコミュニティとなる．

この後，コミュニティ内のノードとの次数が大きなコミュニティ内のノードを選び，新たにシードノードに加えて手順 2〜9 を繰り返す．終了条件はコミュニティの状態が収束するまで，または所望の規模のコミュニティが抽出できたときとする．

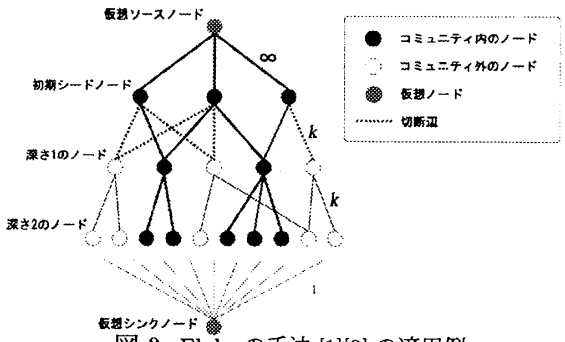

図 3: Flake の手法 [1][2] の適用例

4. 提案手法

Flake の手法では初期シードノードを複数個としている．この場合，初期シードノード同士が異なるコミュニティに属していると，本来取り出されるべきコミュニティが抽出されずに，明らかにコミュニティとはみなせない，複数のコミュニティが混在したノード集合が抽出されてしまうと考えられる．そこで既存手法の初期シードノード数を 1 つとするコミュニティ抽出法を提案する．初期シードノードを 1 つにすれば，そのノードが属しているコミュニティが抽出される性質を利用して，いくつものコミュニティが混ざり合うことを回避することができる．

図 4 に単独シードノードによる最大フローアルゴリズムを用いた提案手法の手順を示す．

手順 1 で辺容量 k を設定し，手順 2 で周辺グラフの探索をする．周辺グラフは，シードノードとリンク 2 以内の部分グラフである．シードノードとリンク 1 のノードを深さ 1 のノード，リンク 2 のノードを深さ 2 のノードと呼ぶ．手順 3 では，ノード間に双方向の辺を設けている．手順 4〜6 では，仮想ソースノードと仮想シンクノードを追加して，仮想シンクノードとシードノードの間，仮想シンクノードと周辺グラフのノードの間に辺を設けている．さらに，仮想ソースノードに接続しているすべての辺は仮想ソースノードを始点とし，仮想シンクノードと接続しているすべての辺は仮想シンクノードを終点としている．このとき，仮想シンクノードとシードノード間の辺容量は ∞ で，仮想シンクノードと周辺グラフのノード間の辺容量は 1 である．手順 7 から最大フローが求まり，手順 8 から最小カットが得られる．得られた最小カットにより切断されたノード集合がコミュニティとして抽出される．

図 5 は，手順 1〜8 を適用後のグラフである．図 5 でも図 3 と同様に有向辺を省略している．手順 8 終了後，コミュニティ内のノードと最も多く隣接している次数が大きなコミュニティ内のノードを選んで，シードノードに追加し手順 2〜9 を抽出されるノード集合 Com の要素が一定になるまで繰り返す．

入力 初期シードノード集合 $S = \{v_1^s\}$
出力 コミュニティ $Com = \{v_1^s, \ldots, v_i^s\}(i = 1, \ldots, n)$

1. k を任意の数にする．
   $n = 1$ とする．
2. $G = (V, E)$；各 $v_i^s \in S$ から深さ 2 以内の周辺グラフ．
3. すべての $(u, v) \in E$ において辺容量 $c(u, v) = k$ とする．
   $(u, v) \in E$，かつ，$(v, u) \notin E$ のとき，辺容量 $c(v, u) = k$ の辺 (v, u) を E に加える．
4. 仮想ソースノード s，仮想シンクノード t を V に加える．
5. s から各 $v_i^s \in S(i = 1, \ldots, n)$ への辺 (s, v_i^s) を E に加える． 辺容量 $c(s, v_i^s) = \infty$ とする．
6. すべての $v \in V$(但し，$v \notin S \cup \{s, t\}$) について辺容量 $c(v, t) = 1$ の辺 (v, t) を E に加える．
7. s-t 最大フローアルゴリズムを実行する．
8. シードノードから不飽和辺をたどって到達可能なノード集合をコミュニティ Com とする．
9. $n \leftarrow n + 1$ とする．
   Com 内の各ノードを次数に基づいてランク付け，このうち高いランクのノードを v_n^s として S に加える．
10. 手順 2〜9 を抽出されるノード集合 Com の要素が一定になるまで繰り返す．

図 4: 単独初期シードノードによる最大フローアルゴリズムを用いた提案手法

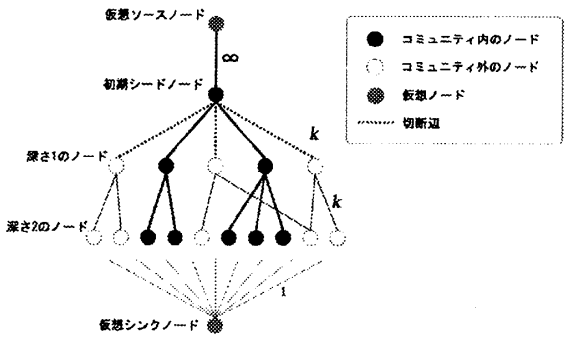

図 5: 提案手法の適用例

既存手法では初期シードノード数を辺容量 k と同一としている．しかし，提案手法においては初期シードノード数が 1 であるので辺容量 $k=1$ として固定すると，初期シードノードのみがコミュニティとして抽出されてしまう [1][2]．そこで提案手法では辺容量 k の値を変化させて適用する．

5. 実験方法および考察

5.1 実験方法

アルゴリズムの評価には，2 章で示したコミュニティの定義,「コミュニティとみなされる部分グラフにおいては，部分グラフ内のノード同士の繋がりが外のノードとの繋がりよりも強い」，を明らかにみたす部分グラフを含む図 6 のグラフモデルを用いる．このグラフモデルは K_6，K_8，K_{10} の 3 つの完全グラフを一本の辺で連結させたグラフであり，K_6，K_8，K_{10} の各部分グラフは，上述のコミュニティの定義をみたしている．図 6 では，コミュニティ別にノードを色分けし，黒は K_6，白は K_8，灰は K_{10} に属しているノードとする．

図 6 のノード No.2, 3, 6, 13, 14, 23 は 2 つのコミュニティの境界上に位置しており，これらのノードを境界ノードと呼び，境界ノード以外のノードを内部ノードと呼ぶことにする．

このグラフモデルに既存手法と提案手法を各々適用し，抽出されるコミュニティの妥当性を考察する．具体的には抽出されたコミュニティのクラスター係数を計算し妥当性の尺度とする．実験では，既存手法，提案手法に対して，初期シードノードの選び方を変化させ，提案手法に対しては，辺容量 k の影響を調べる実験を行った．

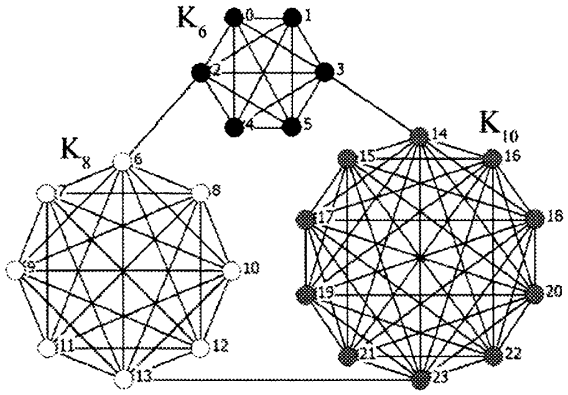

図 6: 複数の完全グラフを含むグラフモデル

5.2 評価方法

図 6 に示したグラフモデルに Flake の手法，提案手法をそれぞれ適用し，抽出される部分グラフのクラスター係数 C を妥当性の尺度とする．クラスター係数 C は，(1) 式に示すようにコミュニティのノード集合のそれぞれのノードのクラスター係数 C_i の平均値である．(1) 式においてノード数を N，ノード i が持つ辺数を k_i，クラスター数を E_i とする．クラスター数 E_i とはノード i が他の 2 つのノードと三角形を構成している組み合わせの数のことである．

$$C = \frac{1}{N}\sum_{i=1}^{N} Ci, \quad C_i = \frac{2E_i}{k_i(k_i-1)} \qquad (1)$$

5.3 実験結果

実験結果を表 1 と表 2 に示す．Flake の手法では，初期シードノードが複数個のため初期シードノードの選び方に応じて，以下のように場合分けができる．

表 1 の 1 列目はシード $*-*$ で，ハイフンの左側の数字は初期シードノードのうち境界ノードであるノードの個数，右側は内部ノードである個数をそれぞれ示している．また，表中の結果は初期シードノードを同じ完全グラフ内のノードの中から選んだ場合のクラスター係数を示している．初期シードノードが K_6，K_8，K_{10} の複数の完全グラフから選ばれている場合は，組み合わせが数多く存在し，この場合はいずれも定義をみたすコミュニティが抽出されなかったため表 1 では割愛した．数値が入っていない箇所は，初期シードノードの組み合わせが存在しない場合である．

提案手法では，初期シードノードが境界ノードの場合と内部ノードの場合で結果が異なるので，表を 2 つに分けている．表 2 中の k は辺容量である．また，灰色に塗られた箇所は，クラスター係数 $C=1$ の箇所である．

表 1: 既存手法の実験結果

シード	K_6	K_8	K_{10}
0-1	0.000	0.000	0.000
0-2	1.000	1.000	1.000
0-3	1.000	1.000	1.000
0-4	1.000	1.000	1.000
0-5	——	1.000	1.000
0-6	——	1.000	1.000
0-7	——	——	0.935
0-8	——	——	0.935
1-0	0.000	0.000	0.000
1-1	1.000	1.000	1.000
1-2	1.000	1.000	1.000
1-3	1.000	1.000	1.000
1-4	1.000	1.000	1.000
1-5	——	1.000	1.000
1-6	——	1.000	1.000
1-7	——	——	0.935
1-8	——	——	0.935
2-0	1.000	1.000	1.000
2-1	1.000	1.000	1.000
2-2	1.000	1.000	1.000
2-3	1.000	1.000	1.000
2-4	1.000	1.000	1.000
2-5	——	1.000	1.000
2-6	——	1.000	1.000
2-7	——	——	0.935
2-8	——	——	0.935

5.4 実験 1 の考察

既存手法

実験結果より，既存手法でコミュニティ抽出を成功させるためには，初期シードノード数がある範囲内に収まっている必要がある．ここで用いた図 6 のグラフモデルに対しては 2 以上の初期シードノード数が必要であり，K_{10} に対しては内部ノードが 6 以下でなければならない

表 2: 提案手法の実験結果
境界ノードが初期シードノードの場合

辺容量 k	K_6	K_8	K_{10}
1	0.000	0.000	0.000
2	1.000	1.000	1.000
3	1.000	1.000	1.000
4	1.000	1.000	1.000
5	1.000	1.000	1.000
6	1.000	1.000	1.000
7	1.000	0.935	0.935
8	0.935	0.935	0.935
9	0.935	0.935	0.935
⋮	⋮	⋮	⋮

内部ノードが初期シードノードの場合

辺容量 k	K_6	K_8	K_{10}
1	0.000	0.000	0.000
2	1.000	1.000	1.000
3	1.000	1.000	1.000
4	1.000	1.000	1.000
5	1.000	1.000	1.000
6	1.000	1.000	1.000
7	1.000	0.935	0.935
8	0.935	0.935	0.935
9	0.935	0.935	0.935
⋮	⋮	⋮	⋮

ことがわかる．すなわち，この条件をみたすような初期シードノード数の調整が予め必要となる．

また，表 1 で割愛した複数の完全グラフのノード同士が混在して初期シードノードになっている場合，クラスター係数 $C = 1$ となるコミュニティは 1 つも抽出されていない．さらに，本稿では紹介していないが，対象を非連結グラフとした場合，定義をまったくみたしていないコミュニティが抽出されてしまう．

今回のグラフモデルでは初期シードノードの組み合わせが $(2^{24} - 1) = 16,777,215$ 通りの組み合わせが考えられる．このうち複数の完全グラフのノード同士が混在して初期シードノードになっている場合は全体の約 99.9% にあたる 16,775,870 通りにのぼる．任意のグラフモデルでは今回のグラフモデル以上に，ノード同士が混在して初期シードノードになる組み合わせが多いと考えられる．このように既存手法では初期シードノードを選択した時点でのコミュニティの混在が一番の問題であるといえる．

提案手法

提案手法では K_6 中のノードが初期シードノードのとき，$2 \leq k \leq 8$ でクラスター係数 $C = 1$ となる．K_8 中のノードが初期シードノードのとき，$2 \leq k \leq 6$ でクラスター係数 $C = 1$ となる．K_{10} 中のノードが初期シードノードのとき，$2 \leq k \leq 6$ でクラスター係数 $C = 1$ となる．

以上のことから，提案手法によってクラスター係数 $C = 1$ となるコミュニティを抽出させる条件として，「辺容量 k の範囲が 2 以上，他完全グラフの境界ノードの次数以下である」という条件が必要になることがわかった．

実験の結果，図 6 のグラフモデルに対しては初期シードノードの選択によらず，この条件を満たしてさえいればコミュニティ抽出に成功することがわかった．また，提案手法では初期シードノードを単一に設定したため，既存手法のように初期シードノードが複数の完全グラフにまたがるという問題は起こらなかった．

6. おわりに

本論文では，Flake の手法が明らかにコミュニティとみなすことのできる完全グラフを抽出できるのかどうかに着目し，アルゴリズムを抽出されたグラフ構造に基づいて評価する方法を提案した．また，これらの評価規範で既存手法を再考し，問題点を改善するアルゴリズムを提案した．

前者については，クラスター係数を導入し評価とした．今回対象にしたのは完全グラフを組み合わせたグラフモデルであったため，クラスター係数が極端に低い値をとなることはなかったが，実験を通じてクラスター係数がある程度大きい妥当なコミュニティを抽出するための辺容量などのパラメータ設定における条件を明らかにした．

後者については，初期シードノードが複数個に既存手法では，非連結なノード集合を含むコミュニティを抽出することが判明した．このコミュニティは明らかに定義をみたしていない．この問題点に対して，初期シードノードを単独とする手法を提案し，この手法では非連結なコミュニティが抽出されるという問題は発生しないことを確かめた．既存手法と提案手法は本研究で用いた評価規範 C においてはほぼ同等であるが，非連結のコミュニティを抽出するか否かが両手法の差である．

今後の課題として，初期シードノードを 1 つとした場合で最大フローアルゴリズムを用いた提案手法が，本論文で用いた完全グラフを連結させたグラフモデルに限らず，任意の構造を持つグラフにおける適用可能性を検証しなければならない．また最大フローアルゴリズムを用いた手法で抽出したコミュニティの規模は比較的サイズが小さいため [6]，大きなコミュニティを抽出する方法を考える必要がある．

参考文献

[1] G.W.Flake, S.Lawrence, C.L.Giles, "Efficient Identification of Web Communities", In 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.150-160, 2000.

[2] G.W.Flake, S.Lawrence, C.L.Giles, F.Coetzee, "Self-Organization and Identification of Web Communities", IEEE Computer, 35(3), pp.66-71, 2002.

[3] J.M.Kleinberg, "Authoritative Sources in a Hyperlinked Environment", In Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, pp.668-677, 1998.

[4] Y.Matsuo, Y.Ohsawa, "Finding the Meaning of Clusters", AAAI Fall Symposium Technical Report FS-02-01, American Association for Artificial Intelligenc, pp.7-13, Cape Cod (2002)

[5] N.Imafuji, M.Kitsuregawa, "Finding Web communities by Maximum Flow Algorithm using Well-Assigned Edge Capacities", To be appeared in Information Processing Technology for Web Utilization, IEICE, 2004.

[6] 今藤紀子，喜連川優，"Max-Flow コミュニティグラフとその特徴分析"，電子情報通信学会 第 15 回データ工学ワークショップ (DEWS2004), 6-B-05, 2004.3