Vol. 49 No. SIG 4(TOM 20) 情報処理学会論文誌：数理モデル化と応用 Mar. 2008

ブログユーザ空間からの重複を許した頻出コミュニティ抽出法

高木 允[†1,†2] 森 康真[†1]
田村慶一[†1] 北上 始[†1]

本研究では，ブログの書き手であるブロガに焦点を当て，ブロガをノード，トラックバックによるつながりを辺としたグラフから，数カ月にわたって頻出し，かつ重複を許したコミュニティを発見する手法を提案する．提案手法は，複数のグラフから頻出部分グラフを抽出し，得られた頻出部分グラフに重複を許したクラスタリング手法を適用することにより，重複を許した頻出コミュニティを発見する．頻出部分グラフの抽出については，頻出部分グラフ抽出の問題を頻出アイテム集合抽出の問題に変換し，LCM 法を用いることで頻出部分グラフ抽出を達成している．重複を許したクラスタリングについては，頻出部分グラフを Newman らのクラスタリング手法を応用し，縮約グラフの作成と再クラスタリングすることで達成している．提案手法の有用性を確認するために，複数カ月にわたりブログデータを収集し，頻出コミュニティの抽出を行った．その結果，共通の興味・関心を持って頻出するコミュニティと，複数のコミュニティに重複してクラスタリングされるブロガを発見できた．

Extraction Method of Overlapping Frequent Communities from Blog User Spaces

Makoto Takaki,[†1,†2] Yasuma Mori,[†1] Keiichi Tamura[†1]
and Hajime Kitakami[†1]

In this study, we focus on bloggers who are writers of blog articles and propose a technique which extracts frequent and overlapped communities across multiple months from graphs consisting of nodes and edges. A node is defined as a blogger and an edge is a connection of trackback. First, the proposed technique extracts frequent communities by extracting frequent subgraphs. Second, the proposed technique extracts overlapping communities by clustering the extracted subgraphs. In the procedures of extraction of frequent subgraphs, we transform the frequent subgraphs extraction problem to the frequent itemsets extraction problem. In the first step, the LCM algorithm is applied to extract the frequent itemsets. In the second step, we applied the Newman's algorithm to find overlapping clusters. To confirm the availability of proposed technique, we collected the graph data and extracted the frequent communities. As a result, frequent communities which have common interests and the bloggers who are clustered into multiple clusters are extracted.

1. はじめに

ウェブログ（ブログ）の登場により，ウェブに関する深い知識を持たない人々も容易に情報を発信できるようになっている．ブログは個人の意見を反映したものが多く，世の中の動きを知るうえでブログ空間から有益な知識を発見することが重要な課題となっている．

ブログ空間からの知識発見に関する研究として，コミュニティ抽出の研究が様々に行われている[1)–3)]．これらの研究はブログ記事を 1 つのノードとし，記事を収集した時点でのスナップショットからのコミュニティ抽出を試みている．

著者らは，ブログ記事ではなく，ブログの書き手であるブロガに着目し，ブロガをノード，記事のトラックバックに基づくブロガ同士のつながりを辺と見なしたグラフ構造に着目している[4)]．その中で，ある一定の期間ごとに発生するグラフの集合を時系列グラフと呼び，その時系列グラフから頻出かつ重複を許したコミュニティを発見することを目標としている．ここで，頻出するコミュニティとは，時系列グラフから抽出される頻出する部分グラフ中に存在するクラスタが特定の話題に偏ったブログ記事を持つコミュニティであると定義する．また，重複を許したコミュニティとは，

†1 広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University
†2 日本学術振興会特別研究員 DC
JSPS Research Fellow

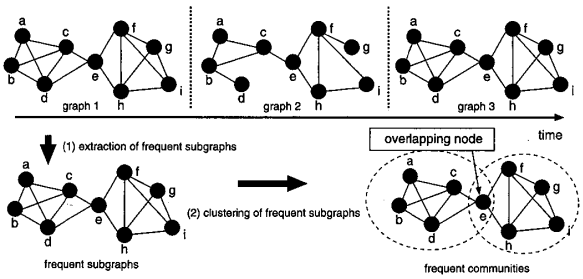


図 1 本研究の概要
Fig. 1 Summary of this study.

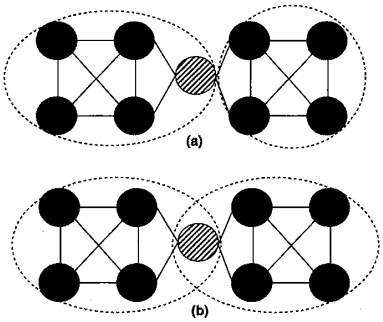


図 2 クラスタリングの種類
Fig. 2 Type of clustering methods.

あるノードが複数のコミュニティに所属することを許したコミュニティであると定義する．

本論文では，前述の時系列グラフから頻出かつ重複を許したコミュニティを発見するために，以下の 2 つの処理を用いた方法を提案する．

(1) 時系列グラフから頻出部分グラフを抽出．
(2) 頻出部分グラフをクラスタリングし，重複を許した頻出コミュニティを抽出．

図 1 に上記 2 つの処理に対応した処理手順を示す．(1) の手順を計算するには，様々な既存手法[5)–8)] があるが，いずれも一般的なグラフの問題を扱っているため，計算時間が大幅に増加するという問題がある．この問題を解決するために，本研究では頻出部分グラフ抽出の問題を頻出アイテム集合抽出の問題に変換し，頻出部分グラフ抽出の高速化を実現する．具体的には，この問題の変換後に文献 9) で提案されている LCM アルゴリズムを用いて頻出アイテム集合を抽出し，その後，逆変換によって，頻出部分グラフを復元する．

(2) では，(1) で得られた頻出部分グラフに Newman らが提案しているクラスタリング手法[10)] (以下，Newman 法) を適用する．Newman 法は，大規模なグラフでも，高速にクラスタリングできる手法として知られている．しかし，Newman 法をそのまま適用すると，クラスタリング結果はクリスプなものとなるという問題がある．つまり，あるノードは必ず 1 つのクラスタにしか所属しない結果となる (**図 2** (a))．人をノードと見なしている本研究においては，より柔軟性を持たせたクラスタリングが望ましい (図 2 (b))．この問題を解決するために，頻出部分グラフへ Newman 法を適用後，縮約グラフを作成し，再度 Newman 法を用いてクラスタリングすることにより，重複を許したクラスタリングを実現する．

抽出されるコミュニティの特徴としては，特定の話題について長期間議論しているコミュニティであり，辺の意味を考慮せずクラスタリングするため，特定の話題でつながっているだけでなく，コミュニティ内で何らかの交流関係や社会的なつながりを持つコミュニティであると考えられる．また，抽出された頻出コミュニティに向けた効率的なマーケティング，特定の話題に特化したブログ検索，ブロガへの情報推薦などへの応用も期待できる．

提案手法の有効性を示すために，収集したデータに提案手法を適用した．結果として，収集したデータから，特定の話題について長期間議論しているコミュニティを発見でき，複数のコミュニティに所属しているブロガを発見することができた．

本論文の構成は以下のとおりである．2 章で関連研究について述べ，3 章で本研究において使用する Newman 法について説明する．4 章で提案手法について説明し，例を用いて頻出部分グラフ抽出とクラスタリングについて説明する．5 章でデータ収集についての説明と，収集したデータに提案手法を適用した結果・考察を示す．最後の 6 章でまとめを行う．

2. 関連研究

本章では，本研究の関連研究として，「コミュニティの発見・解析に関する研究」，「グラフマイニングに関する研究」，「重複を許したクラスタリングに関する研究」の 3 つの視点から述べる．

2.1 コミュニティの発見・解析に関する研究

従来のコミュニティに関する研究[1),11)–15)] では，Web 上の記事やブログ記事をノードとするコミュニティに着目しているのに対し，本研究では人 (ブロガ) をノードとし，時間経過とともに頻繁に現れる人のコミュニティを見つけ出すことに着目している点が大きく異なる．すなわち，本研究では，消滅しやすいコミュニティよりも長期的に安定したつながりを持つ，人のコミュニティの抽出に着目している．

2.2 グラフマイニングに関する研究

与えられたグラフデータベース $D = \{G_1, \cdots, G_n\}$ から頻出する部分グラフを抽出する研究が様々行われている[5)–8)]．文献 5)–7) では，同一ラベルを持つノー

ドや辺が複数存在する一般グラフから同型な頻出部分グラフを抽出する方法が提案されている．一般グラフから，同型な部分グラフを生成するには多大な計算時間を要するため，これらの文献では，各 G_i のノード数が 15 から 40 の疎なグラフを扱う程度にとどめている．

本研究で扱うグラフは，ノードと辺のどちらにも同一ラベルを許さない時系列グラフであるので，一般グラフとは異なる．また，各 G_i のノード数が数百から数千ノード規模であり，6 カ月間から 12 カ月間にわたり収集した時系列グラフを想定している．以上により，文献 5)–7) の手法を我々が想定している頻出かつ重複を許したコミュニティ発見の問題に応用するには，計算時間の面で不向きである．

文献 8) で提案されている CODENSE は，カットを用いた頻出部分グラフ抽出法として知られている．しかし，著者らが収集したデータに応用した実験では，本来 1 つになるべきコミュニティがサイズの小さな頻出部分グラフに分割されるという問題が生じている．また，スター構造のようなコミュニティは，頻出部分グラフとして抽出できないという問題も生じている．

そのため，本研究では，グラフデータの性質を利用し，頻出部分グラフ抽出の問題を頻出アイテム集合抽出の問題に変換し，高速に頻出部分グラフを抽出する手法を提案している．

2.3 重複を許したクラスタリングに関する研究

近年，さかんに研究が行われているのが，ネットワーク構造解析に基づくクラスタリング手法である[16),17)]．ネットワーク構造解析に基づくクラスタリング手法は，クラスタリング後に全体の構造，特に，クラスタとクラスタとのつながり方を取り出せるという利点を持っている．

ネットワーク構造解析に基づくクラスタリング手法は，データクラスタリングにおける類似度などのデータをノードに置き換えたものであるといえる．データクラスタリングでは，fuzzy c-means 法[18),19)] など，fuzzy クラスタリングにおいて重複を許したクラスタリング手法が研究されている[20),21)]．一方，ネットワーク構造解析に基づくクラスタリング手法については，重複を許した手法に関する研究は少ない．

Palla らは，文献 16) において，クリークパーコレーション法と呼ばれるネットワーク構造解析に基づくクラスタリング手法を提案している．クリークパーコレーション法では，ユーザが与えた値である k を基にすべてのクリークを列挙していく．すべてのクリークを列挙しているため，結果的に，複数のクラスタにクラスタリングされる頂点が存在する．しかしながら，k の値をどのように設定するか，また，どの k により最適なクラスタが抽出されるかは，ユーザが得られた結果を分析する必要がある．k の値が小さいと全体が 1 つのクラスタとなり，k の値が大きいと，少数のクリークが抽出されるだけである．また，クリークしか取り出されないため，全体を分類するのには不向きであるという欠点が存在する．

Zhang らは，Graph Spectral Cut と，fuzzy c-means 法を利用した，ネットワーク構造解析に基づくクラスタリング手法を提案している[17)]．提案されている手法では，Graph Spectral Cut の途中段階で最適なカットを見つけるために使用されている k-means 法の部分を fuzzy c-means 法に置き換えたものであり，重複を許すクラスタリング手法となっている．

この手法では，fuzzy c-means 法の実行後，クラスタへの所属比率が λ 以上のノードを，複数のクラスタにクラスタリングするという閾値 λ の設定が必要である．λ の設定は，ネットワークの規模や構造に依存するため，最適な λ の値を何らかの手法で見つける必要があり，最適な λ を見つけるための指導原理的な指標が存在しない．

そのため，本研究においては，パラメータ設定の必要がないネットワーク構造解析に基づいた，Newman 法を応用する方法を考えた．

3. Newman 法

本章では，クラスタリングの際に用いる Newman 法[10)] について説明する．

無向グラフ $G(V,E)$ が与えられ，隣接行列 A の要素 A_{vw} が以下のように与えられているとする．

$$A_{vw} = \begin{cases} 1 \text{（頂点 } v \text{ と頂点 } w \text{ がつながっている場合）} \\ 0 \text{ （その他の場合）} \end{cases}$$

頂点 v の次数 k_v は以下の式で表される．

$$k_v = \sum_{w \in V} A_{vw}$$

頂点 v が所属するクラスタ番号を c_v とし，頂点 v と頂点 w とが同じクラスタに所属するかどうかを示す関数を以下のように，

$$\delta(c_v, c_w) = \begin{cases} 1 \ (c_v = c_w \text{ の場合}) \\ 0 \ (\text{その他の場合}) \end{cases}$$

と定義する．

クラスタ内部に存在する辺の割合が多いクラスタリング結果ほど，良いクラスタリングといえる．ここで，

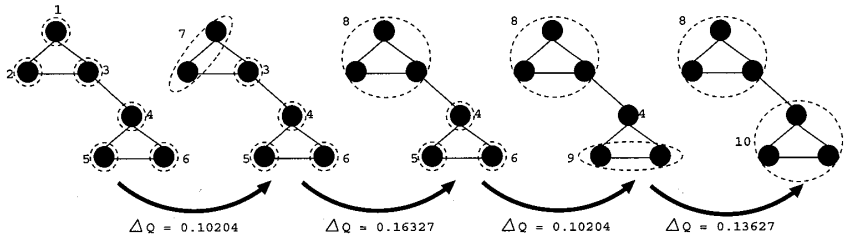


図 3 Newman 法の概要
Fig. 3 Summary of Newman's algorithm.

クラスタ内部に存在する辺の割合を式で表すと，

$$\frac{\sum_{v,w\in V} A_{vw}\delta(c_v,c_w)}{\sum_{v,w\in V} A_{vw}} = \frac{1}{2m}\sum_{v,w\in V} A_{vw}\,\delta(c_v,c_w) \qquad (1)$$

となる．ここで，m はグラフ中の辺の総数である．式 (1) の値が大きなクラスタリングほど内部に存在する辺の割合が多くなり，評価の高いクラスタリングであるといえる．

しかしながら，この式で表される割合だけでは，グラフ全体を 1 つのクラスタとするとき最大値 1 となっているために，評価値としては使用できない．そこで，辺をランダムに張り替えたとき，頂点 v と頂点 w の間に辺が張られる確率 $k_vk_w/2m$ を A_{vw} から引いた値を A_{vw} と置き換えたものをモジュール性の度合い Q として定義する．

$$Q = \frac{1}{2m}\sum_{v,w\in V}\left[A_{vw} - \frac{k_vk_w}{2m}\right]\delta(c_v,c_w) \qquad (2)$$

ここで，

$$e_{ij} = \frac{1}{2m}\sum_{v,w\in V} A_{vw}\delta(c_v,i)\delta(c_w,j),$$
$$a_i = \frac{1}{2m}\sum_{v,w\in V} k_v\delta(c_v,i)$$

とおくと，Q は以下のようになる．

$$Q = \sum_i (e_{ii} - a_i^2) \qquad (3)$$

Newman 法は，Q の値を最大にするようなクラスタリング結果を求める組合せ最適化問題となる．組合せの数は頂点数の指数オーダ存在するため，厳密解を求めるのではなく，貪欲アルゴリズムにより，近似最適解を求めている．最初に 1 つの頂点を 1 つのクラスタとし，階層的にクラスタどうしを結合する．どのクラスタどうしを結合するかは，2 つのクラスタを結合することにより，Q の値がどれだけ増加するかで判断する．

クラスタ i とクラスタ j とを結合したときに，増加する Q の値である ΔQ_{ij} は以下の式により求めることができる．

$$\Delta Q_{ij} = 2(e_{ij} - a_ia_j) \qquad (4)$$

図 3 に，Newman 法のアルゴリズムの概要を示す．最初に，グラフの各頂点を 1 つのクラスタとする．次に，2 つのクラスタ i とクラスタ j を結合したときのモジュール性の増加値 ΔQ_{ij} を計算する．図 3 では，ΔQ_{12} の値が最も高いため，最初にクラスタ 1 とクラスタ 2 が結合されている．クラスタ 1 とクラスタ 2 を結合してできたクラスタに新たなクラスタ番号を振り，結合を続けていく．最終的に，クラスタ 8 とクラスタ 10 の 2 つのクラスタが残り，クラスタリングの処理を終了し，クラスタ 8 とクラスタ 10 が結果として抽出される．

4. 提案手法

本章では，提案手法のアルゴリズムと，具体的な例を用いた説明を，頻出部分グラフの抽出と頻出部分グラフから重複を許したクラスタの抽出に分けて行う．

4.1 アルゴリズム

ブロガをノード，トラックバックによるつながりを辺とした重みなし・無向単純グラフの集合であるグラフデータベースを $D = \{G_1, \cdots, G_n\}$ とする．以後，u, v は 1 つのノード，$\{u, v\}$ は辺を表す記号とする．D 中のグラフ G_i は $G_i = G(V_i, E_i)$，$E_i \subset \{\{u,v\}|u, v \in V_i, u \neq v\}$，$\{u,v\} \in E_i$，$1 \leq i \leq n$ と定義される．V_i はノードの集合，E_i は辺の集合である．**図 4** (a) に D の例を示す．提案手法のアルゴリズムは大きく以下の 2 つのステップからなる．

(1) 頻出部分グラフの抽出
(2) 頻出部分グラフから頻出コミュニティとなりうる重複を許したクラスタの抽出

以下の節でそれぞれのステップの詳細について説明する．

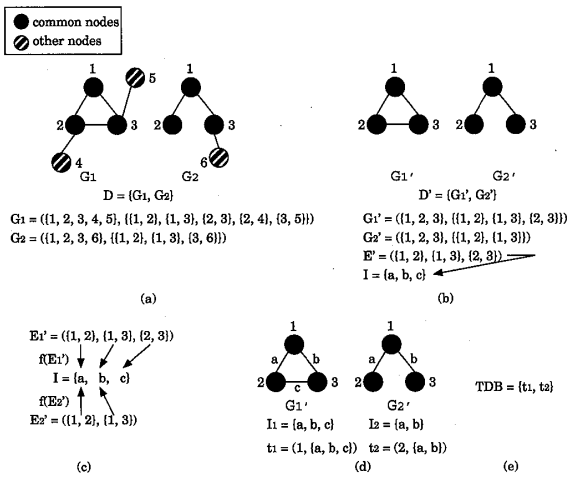


図 4 トランザクションデータベース作成例
Fig. 4 An example of creation of TDB.

4.2 頻出部分グラフの抽出

頻出部分グラフの抽出においては，頻出部分グラフ抽出の問題を頻出アイテム集合抽出の問題に変換し，得られた頻出アイテム集合を逆変換することで，頻出部分グラフを得る．頻出部分グラフ抽出アルゴリズムを，Algorithm 1 に，具体的な例を用いた説明を図 4 に示す．頻出部分グラフを抽出するための前処理として，$D = \{G_1, \cdots, G_n\}$ からすべての G_i について共通しているノードを取り出したグラフデータベース $D' = \{G'_1, \cdots, G'_n\}$ を作成する．図 4 (a) 中の塗りつぶされているノードをすべての G_i に共通しているノードだとすると，図 4 (a) 中の斜線で示されているノードを除去することになる．前処理を実行すると，図 4 (b) の状態になる．

$G'_i = G(V', E'_i)$ と定義すると，ノード集合 V' はすべての G'_i に対して同一である．つまり，$V' = V_1 \cap V_2 \cap \cdots \cap V_n$ である．また $E'_i = \{\{u', v'\} | u', v' \in V', \{u', v'\} \in E_i\}$ と定義できる．E'_i は V' に含まれるノードのペアのみで構成されている．

次に，問題変換のために，グラフの辺にラベルを付与する．E' をすべての E'_i の和集合とする．$|E'|$ 個のラベルを要素としたラベル集合 I を作成する．E' から I への全単射を f と定義する（図 4 (c)）．

全単射 f を用いて E'_i からラベル集合 $I_i = \{label_{i1}, \cdots, label_{i|E'_i|}\}$ を作成する（図 4 (d)）．このラベル集合をアイテム集合と見なし，トランザクションデータベース $TDB = \{t_1, \cdots, t_n\}$ を作成する（図 4 (e)）．ここで，$t_i = (i, I_i)$ である．

本研究では，TDB から極大頻出アイテム集合を抽出するために，文献 9) で提案されている LCM 法を用いた．LCM 法は本来，頻出アイテム集合を高速に求めるために提案されている手法であるが，問題の変換により，本研究で扱っているデータに適用することが可能となる．様々な頻出アイテム集合に関する研究が行われているが，他の手法では，すべての頻出アイテム集合が抽出され，冗長かつ高速性に欠ける．そこで，大量のアイテム集合から高速に極大頻出アイテム集合のみを抽出することができる，LCM 法を用いた．ここでは，極大頻出アイテム集合を求めているが，包含関係にある小さな頻出部分グラフなどの抽出を避けるためである．

Algorithm 1 EXTRACT_FR_SUBG(min_sup,D')

```
1: TDB := φ; MAX_PAT := φ;
2: E' := φ; FR_SUBG := φ;
3: for all E'_i ∈ G'_i do
4:     E' := E' ∪ E'_i;  /*Creating Edge Set*/
5: end for
6: Create Label Set I;  /*|I| = |E'|*/
7: Define f as a bijection from E' to I;
8: for all E'_i ∈ G'_i do
9:     t_i := φ; I_i := φ;
10:    I_i := f(E'_i);  /*Creating Itemset*/
11:    t_i := (i, I_i);  /*Add i and I_i to t_i*/
12:    TDB := TDB ∪ t_i;  /* Add Transaction
       to TDB*/
13: end for
    /* Extract Maximal Frequent Patterns */
14: MAX_PAT := EX_MAX(min_sup, TDB);
    /* Mapping Itemsets to Edges */
15: for all PAT_i ∈ MAX_PAT do
16:    FSG_i := φ; FE_i := φ; FV_i := φ;
17:    FE_i := f^{-1}(PAT_i); Create FV_i from FE_i;
18:    FSG_i := G(FE_i, FV_i);
19:    FR_SUBG := FR_SUBG ∪ FSG_i;
20: end for
21: return FR_SUBG;
```

最小支持数を min_sup とし，極大頻出アイテム集合を抽出する関数を $EX\_MAX(min\_sup, TDB)$ とする．EX_MAX で得られた極大頻出アイテム集合を $MAX\_PAT = \{PAT_1, \cdots, PAT_l\}$ とする．ここで，$PAT_i = \{label_{i1}, \cdots, label_{im}\}$，$1 \leq i \leq l$，$1 \leq m$ である．PAT_i から，辺の集合 FE_i へ，f の逆写像 f^{-1} を用いてアイテム集合から辺集合へ変換する．

辺集合からノード集合を得て，得られた辺集合とノード集合から頻出部分グラフ FSG を復元する．頻出部分

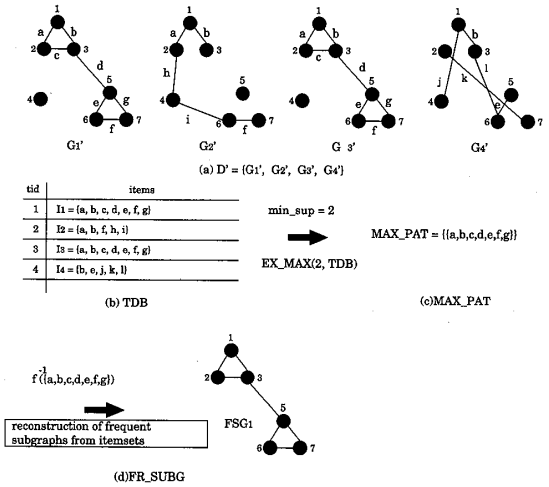


図 5 頻出部分グラフ抽出例
Fig. 5 An example of extraction of frequent subgraphs.

グラフ集合を，$FR\_SUBG = \{FSG_1, \cdots, FSG_l\}$ と定義する．$FSG_i = G(FV_i, FE_i)$，$FE_i = \{\{u_i, v_i\} | u_i, v_i \in V'\}$，$1 \leq i \leq l$ であり，FV_i は FE_i を構成するすべてのノードの集合である．

上記手法により，頻出部分グラフを得る．

4.3 頻出部分グラフ抽出例

本節では，例を用いて**図 5** (a) に示すグラフデータベース $D' = \{G'_1, G'_2, G'_3, G'_4\}$ からの頻出部分グラフ抽出例を示す．

まず，すべての E'_i の和集合 E' を求める．$|E'|$ 個のラベルを持ったラベル集合 I を作成すると，$I = \{a, b, c, d, e, f, g, h, i, j, k, l\}$ を得る．f を用いて，E' をラベル付けする．そして，f を用いて E'_i に対応するラベルをアイテムとしたトランザクション I_i を生成する．さらに，トランザクションデータベース TDB を作成する（図 5 (b)）．このとき，$TDB = \{t_1, t_2, t_3, t_4\}$ であり，$t_1 = (1, I_1)$，$t_2 = (2, I_2)$，$t_3 = (3, I_3)$，$t_4 = (4, I_4)$ である．ただし，$I_1 = \{a, b, c, d, e, f, g\}$，$I_2 = \{a, b, f, h, i\}$，$I_3 = \{a, b, c, d, e, f, g\}$，$I_4 = \{b, e, j, k, l\}$ である．

そして，図 5 (c) に示すように，作成した TDB から，関数 EX_MAX を用いて極大頻出アイテム集合を抽出する．ここで，最小支持数は 2 としている．$EX\_MAX(2, TDB)$ によって得られた極大頻出アイテム集合は $MAX\_PAT = \{PAT_1\}$，$PAT_1 = \{a, b, c, d, e, f, g\}$ である．

抽出された MAX_PAT を，f の逆写像を用いて辺集合へ変換し，変換された辺から元のグラフを復元する（図 5 (d)）．復元されたグラフは $FR\_SUBG = \{FSG_1\}$ となり，$FSG_1 = G(FV_1, FE_1)$，$FV_1 = \{1, 2, 3, 5, 6, 7\}$，$FE_1 = \{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{3, 5\}, \{5, 6\}, \{6, 7\}, \{5, 7\}\}$ である．

4.4 重複を許したクラスタリング手法

本章では，提案手法の 1 つである，Newman 法を応用した，重複を許したクラスタリング手法の説明を行う．提案するクラスタリング手法は，Newman 法でクラスタリングした後，縮約グラフを作成し，再度 Newman 法でクラスタリングを行い，重複してクラスタリングされるノードを見つけ出す．以下に，アルゴリズムの説明，具体的な例を用いた説明を行う．また，3 章で説明した Newman 法を関数 $Newman(G)$ と表記する．

Algorithm 2 Overlapping_Newman 入力：G(V,E) 出力：C

```
1: C := φ; /*最終的な解を保存する変数*/
2: NC := φ; /*Newman(G) によって得られるクラスタリング結果*/
3: NC := Newman(G); /*Newman(G) の実行*/
4: for all NC_i ∈ NC do
5:   {SG_i, v_i} := MAKE_C_GRAPH(G, NC_i) /*縮約グラフの作成*/
6:   DUP_i := FIND_OVERLAP(SG_i, v_i) /*重複するノードを見つける*/
7:   C := C ∪ {NC_i ∪ DUP_i} /*重複するノードを NC_i に追加*/
8: end for
9: return C; /*クラスタリング結果を返す*/
```

Algorithm 2 に，重複を許したクラスタリング手法のアルゴリズムを示す．まず，$Newman(G)$ を用い，入力されたグラフ G のクラスタリングを行う．得られたそれぞれのクラスタを変数 NC に代入する．$Newman(G)$ により得られたクラスタの数が n 個あれば，$NC = \{NC_1, \cdots, NC_n\}$ となる．そして，それぞれの NC_i について，関数 $MAKE\_C\_GRAPH(G, NC_i)$ を用いて，NC_i を 1 つのノードとした縮約グラフを作成する．このとき，1 つのノードに縮約されたクラスタにつながっている辺の数の情報はそのまま保存しておく．$MAKE\_C\_GRAPH(G, NC_i)$ の出力として，縮約グラフ SG_i と縮約されたクラスタであるノード v_i が得られる．

さらに，関数 $FIND\_OVERLAP(SG_i, v_i)$ を用いて，重複してクラスタリングされるノードを見つけ出す．$FIND\_OVERLAP(SG_i, v_i)$ の出力として，重複してクラスタリングされるノード集合 DUP_i が得

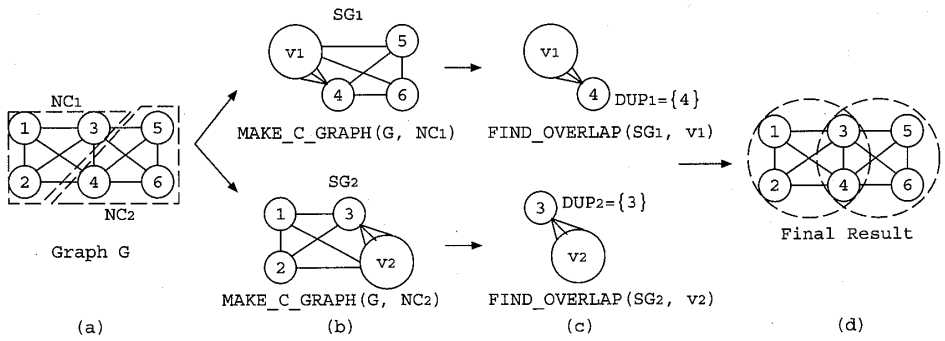


図 6 重複を許したクラスタリングの例
Fig. 6 An example of overlap clustering.

られる．得られたノード集合を NC_i に付加し，最終的な出力である C に挿入する．この手順をすべての NC_i について行い，重複してクラスタリングされるすべてのノードを見つけ出す．

関数 $FIND\_OVERLAP(SG_i, v_i)$ の動作は，次のとおりである．縮約したノード v_i に隣接するノードをそれぞれ v_i と結合したときの ΔQ を計算する．そして，結合後に ΔQ が負となる v_i に隣接しているノードを取り除く．v_i と結合しても ΔQ が負とならないノード集合を結果として出力する．

Newman 法はヒューリスティックな手法であり，厳密な解を求めているわけではない．そのため，1 度識別されたクラスタへ他のクラスタに識別されているノードを追加すると，Q の値が増加する可能性がある．縮約グラフの作成を行い，再度 ΔQ の値を計算することで，1 度目のクラスタリングでは活用されていなかった辺の情報を活用でき，重複してクラスタリングされるノードを見つけ出すことができる．

提案手法を適用すると，最初に Newman 法を適用して識別されたクラスタの境界に存在しているノードしか重複してクラスタリングされるノードの対象にならない．しかしながら，本研究では，クラスタとクラスタの間の橋渡しを行う仲介者や，情報発信者などのブロガを発見することを前提としている．そのため，提案手法を用いクラスタの境界面に存在するノードを対象として重複を取り出している．

4.5 重複を許したクラスタリングの例

本節では，具体的な図を用いて提案手法の説明を行う．**図 6** に提案手法の具体例を示す．6 つのノードで構成されるグラフ G を $Newman(G)$ によりクラスタリングすると，図 6(a) に示すように，NC_1 と NC_2 にクラスタリングされる．次に，$MAKE\_C\_GRAPH(G, NC_i)$ を用いて縮約グラフを作成する．すると，図 6(b) に示す 2 つの縮約グラフ SG_1 と SG_2 が作成される．

次に，$FIND\_OVERLAP(SG_i, v_i)$ を用いて，作成された縮約グラフから重複してクラスタリングされるノードを発見する．図 6(c) に示すように，v_1 とノード 4，v_2 とノード 3 を結合すると ΔQ の値が増加するため，それぞれ重複してクラスタリングされるノード DUP_1，DUP_2 として識別される．

最後に，v_1 と v_2 を元のグラフに復元し，重複を許したクラスタリングを終了する．最終的に，図 6(d) に示すようにノード 3，4 が 2 つのクラスタに重複してクラスタリングされるノードとなる．

5. 評価実験

本章では，まず，データ収集とグラフ作成について説明し，評価実験を行った結果を示す．

5.1 データ収集とグラフ作成

本研究では，**表 1** に示されているホスティングサービスを利用しているブログ記事の収集を行った．収集を行ったブログの記事は 2006 年 1 月 1 日から 2006 年 1 月 31 日までの記事のように，1 カ月単位で，2006 年 8 月まで収集した．つまり，時系列グラフデータは，$D = \{G_1, G_2, G_3, G_4, G_5, G_6, G_7, G_8\}$ となる．各 G_i の詳細を**表 2** に示す．**図 7** にデータ収集方法とグラフ作成についての説明を示す．

収集間隔を 1 カ月ごとにしたのは，収集期間を 1 週間ごとや 2 週間ごとにした場合，収集できるノードの数が大幅に減少してしまうという問題があるためである．過去の記事などへトラックバックを張っていても時間の制約が強くなってしまうため，記事の収集が早い段階で終わってしまう．よって，1 カ月ごとに記事を収集すれば，同月内の過去の記事へのトラックバックもたどることができ，収集できるデータ量が増加するため，1 カ月ごとに記事の収集を行った．

データ収集開始時において，始点となる記事（ブロガ）を選択する．各月ごとに話題が重ならないように始点となる記事をランダムに選択した．始点となる記

表 1 データ収集の対象としたホスティングサービス
Table 1 Hosting services.

FC2 ブログ	ドリコムブログ	goo ブログ
ココログ	livedoor ブログ	me ブログ
Yahoo!ブログ	NetLaputa Blog	アメブロ
Block Blog	So-net blog	JUGEM
楽天ブログ	Yaplog!	Seesaa ブログ
ウェブリブログ	Dream Blog	エキサイトブログ
カフェブロ	Fruit Blog	

表 2 収集したデータの詳細
Table 2 Details of collected data.

グラフ G_i	$\|V_i\|$	$\|E_i\|$	総記事数
G_1（2006 年 1 月）	5,861	28,518	8,383
G_2（2006 年 2 月）	4,699	24,695	8,884
G_3（2006 年 3 月）	4,843	27,091	8,631
G_4（2006 年 4 月）	1,699	9,517	6,143
G_5（2006 年 5 月）	4,010	26,448	12,047
G_6（2006 年 6 月）	4,432	28,733	15,569
G_7（2006 年 7 月）	3,147	18,986	9,878
G_8（2006 年 8 月）	3,951	23,966	9,715

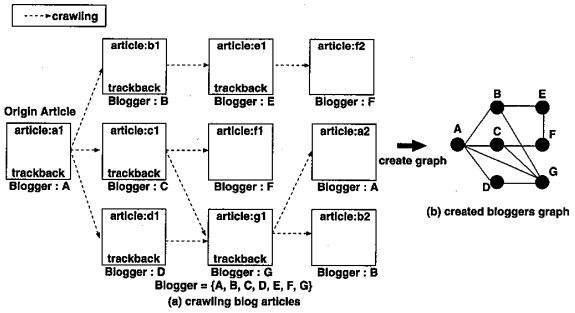


図 7 データ収集方法とグラフの作成
Fig. 7 Data collection and graph creation.

事からトラックバックを抽出し，抽出したトラックバックをたどることにより記事を収集する（図 7 (a)）．この作業をトラックバックがなくなるまで行う．収集先の記事の生成日時が収集範囲外の日時ならばその記事は収集しない．すでに収集済みのブロガが再度トラックバックされていた場合，新たなノードは作成せず，辺のつながりを付け加えその記事を収集する．図 7 (b) のように，データ収集が終了した時点でできあがるグラフのノード数は収集したブロガ数と等しいことになる．各月において他の月に収集したグラフのデータはまったく参照しない．つまり，6 月においてあるノード u，v 間に形成されていた辺があったとしても，7 月のデータを収集する際にはすべてのノードと辺の情報は削除されており，新規にグラフを作成していく．

5.2 実　　験

前節で説明した時系列グラフ D に提案手法を適用した．前処理を行ったところ，すべての G_i に共通して出現しているノードは 172 ノードであった．各 G'_i の詳細を**表 3** に示す．今回の実験では，頻出部分グラフ抽出のための最小支持数を最大の 8 に設定し，頻出部分グラフの抽出を行った．

表 3 各 G'_i の詳細
Table 3 Details of each G'_i.

グラフ G_i	$\|V'_i\|$	$\|E'_i\|$
G'_1（2006 年 1 月）	172	709
G'_2（2006 年 2 月）	172	1082
G'_3（2006 年 3 月）	172	1846
G'_4（2006 年 4 月）	172	844
G'_5（2006 年 5 月）	172	882
G'_6（2006 年 6 月）	172	1340
G'_7（2006 年 7 月）	172	1193
G'_8（2006 年 8 月）	172	1113

時系列グラフ D に提案手法を適用したところ，1 つの頻出部分グラフが抽出され，Newman 法を用いてクラスタリングを行った結果，11 個のクラスタが識別された．識別されたクラスタを可視化したものを**図 8** に示す．図中の 1 から 11 の番号は任意に付与したクラスタ番号を示している．

5.2.1 コミュニティに関する評価

識別された 11 個のクラスタが頻出コミュニティとなりうるのかを調査するために，各クラスタ内のブロガが書いたブログ記事に対して，*tf-idf* 法を用いて，重要キーワードの抽出を行った．各クラスタの重要キーワード上位 5 件とそれぞれのキーワードについての *tf-idf* 法による評価値を**表 4** に示す．

表 4 中には，「ブログ」というキーワードが上位にランキングされている．これは，*tf-idf* 法による解析の際に，ブログ記事の本文以外の文章（コメント，広告など）もノイズとして解析結果の対象となったことが原因である．しかしながら，ここでは，各クラスタにおいて特徴的なキーワードが抽出されていることを考慮し，「ブログ」などの一般的なキーワードをノイズと見なす．

表 4 から，各クラスタとも野球，サッカーの話題に偏っていることが分かる．実際に，手作業で各クラスタ内のブロガの記事を調べてみたところ，*tf-idf* 法の値が高い上位のキーワードに関する記事を主に扱っているブロガが各クラスタ内に多数存在していた．例として，クラスタ 3 にクラスタリングされているブロガが書いているブログ記事を調べたところ，すべてのブロガが広島東洋カープの話題を中心とした記事を書いていた．タイトルや自己紹介などの欄にも自らがカープファンであるということを明記しているブロガがほとんどであった．このことは，*tf-idf* 法による評価値からも分かる．上位にランキングされているキーワー

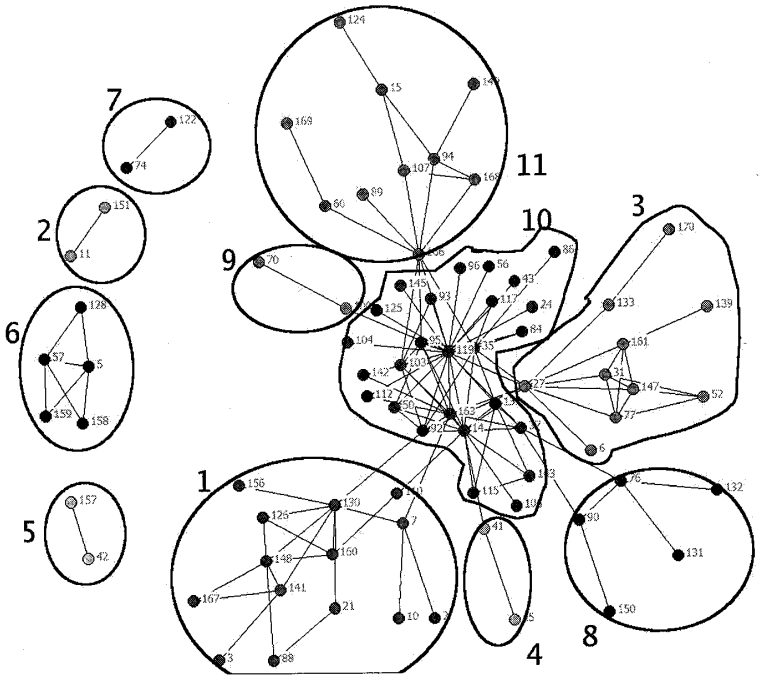


図 8 抽出された頻出コミュニティ
Fig. 8 Extracted frequent communities.

表 4 各クラスタにおける *tf-idf* 法による解析結果（表中の括弧内は *tf-idf* 法により算出された評価値を示す）
Table 4 Results of analysis for each cluster by *tf-idf*.

クラスタ＼ランク	1 位	2 位	3 位	4 位	5 位
1	ホークス (4935)	野球 (4685)	ブログ (4167)	鷹 (3561)	阪神 (2767)
2	マリーンズ (1923)	トラック (1493)	ブログ (1124)	ロッテ (922)	千葉 (912)
3	カープ (6868)	広島 (6207)	鯉 (4237)	永 (4073)	対 (3965)
4	ライオンズ (205)	野球 (190)	イタリア (188)	トラック (186)	ブログ (173)
5	ライオンズ (557)	ステージ (345)	ブログ (327)	北海道 (230)	涌井 (176)
6	野球 (1245)	ライオンズ (1135)	ブログ (901)	競馬 (853)	記事 (813)
7	ブログ (266)	スワローズ (249)	東京 (237)	愛 (232)	記事 (186)
8	日本 (4195)	ドイツ (4061)	サッカー (3793)	イタリア (3736)	決勝 (2494)
9	ブログ (2725)	阪神 (1569)	日記 (978)	トラック (957)	小松 (896)
10	日本 (24728)	ブログ (16631)	ドイツ (15986)	決勝 (15794)	サッカー (14469)
11	中日 (5191)	ドラゴンズ (4399)	日本 (3225)	野球 (2773)	阪神 (2471)

ドの評価値と下位にランキングされているキーワードの評価値には大きな差があり，上位のキーワードがそのクラスタの特徴を表していると考えられる．

クラスタ 3 にクラスタリングされている個々のブロガの記事を *tf-idf* 法を用いて解析した結果を**表 5** に示す．表 5 では，それぞれのブロガについて，*tf-idf* の値の高い，上位 3 件のキーワードを示している．表中のブロガ ID は，著者らが任意に割り当てた識別用の ID を示している．表 5 からも，「広島」，「カープ」，「野球」などのキーワードが上位にランキングされており，クラスタ 3 内の各ブロガが書いている記事は，広島東洋カープのことについて書いていることが分かる．この事実は，表 4 のクラスタ 3 に示すキーワードと強い相関を持っている．

今回の評価実験では，最小支持数を 8 と設定しているため，8 カ月にわたってトラックバックを張り合う関係にあることが分かり，一時的な興味などによるつながりではないことが分かる．以上の事実より，抽出された頻出部分グラフをクラスタリングすることによって得られたクラスタは，頻出コミュニティであるということが可能である．

5.2.2 重複を許したクラスタリングに関する評価

次に，重複してクラスタリングされたノードについて説明する．今回の実験では，ノード番号 27 がクラスタ 3 とクラスタ 10 に重複して所属しているという結果となった．**図 9** に，Newman 法によるクラスタリング結果を示し，**図 10** に提案手法によるクラスタリング結果を示す．図 9 では，単純に Newman 法を

表 5 クラスタ 3 における各ブロガの重要キーワード上位 3 件
Table 5 Important key words of cluster 3.

ブロガ ID ＼ ランク	1 位	2 位	3 位
6	アウト	広島	ランナー
27	広島	カープ	ブログ
31	鯉	カープ	広島
52	カープ	広島	鯉
77	対	公式	カープ
133	対	カープ	広島
139	広島	カープ	中日
147	カープ	鯉	広島
161	カープ	ー	日本
170	永	相互	野球

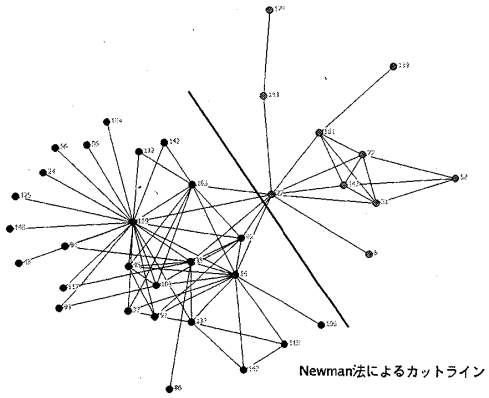


図 9 Newman 法によるクラスタリング結果
Fig. 9 Result of Newman's algorithm.

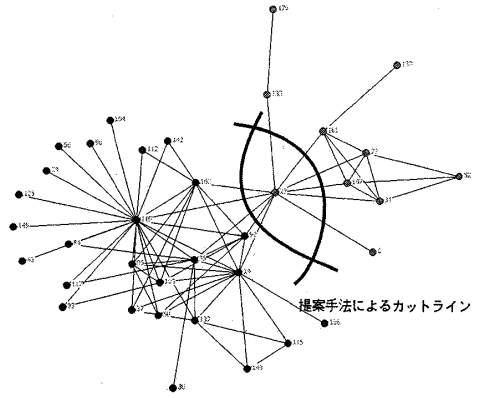


図 10 提案手法によるクラスタリング結果
Fig. 10 Result of proposed method.

使用した場合にクラスタリングされる境界線を示している．提案手法を用いると，図 10 に示すように，ノード 27 がクラスタ 3 とクラスタ 10 に重複してクラスタリングされている．

ここで，ノード 27 のブロガがどのような話題を主に扱っているのかを調べるため，ノード 27 のブロガが保持している記事から，*tf-idf* 法を用いて重要なキーワードの抽出を行った．表 5 には，上位 3 件しか示していないが，ここでは上位 10 件を**表 6** に示す．表 6 から，日本のプロ野球球団である「広島東洋カープ」についてのキーワードが上位に位置していることが分かる．表 6 には表れていないが，上位 30 件までのキーワードを抽出したところ，2006 年に開催されたサッカーワールドカップに関するキーワードが抽出されていた．

表 6 ノード番号 27 の *tf-idf* 法による解析結果
Table 6 Analyzing result of node 27 by *tf-idf*.

ランク	キーワード
1 位	広島
2 位	カープ
3 位	ブログ
4 位	野球
5 位	対
6 位	中日
7 位	パ・リーグ
8 位	日本ハム
9 位	イタリア
10 位	日記

ノード 27 のブロガが書いているブログ記事を手作業で調査したところ，自らが広島東洋カープのファンであると言及しており，主に広島東洋カープについての記事を扱っていること，2006 年 6 月に開催されたサッカーワールドカップについての記事も記述していることが分かった．特に，クラスタ 10 にはその日に開催されたスポーツの結果などを扱っているブロガが多数存在しており，クラスタ 10 にクラスタリングされたブロガとトラックバックを張り合うことで交流を深めており，コミュニティとコミュニティの橋渡しをしている重要なブロガであると考えられる．

重複を許したクラスタリングにより，ノードが複数のクラスタに属することが可能となることで，重複してクラスタリングされるノードの特徴や，重要ノードの特定などが容易になる．

今回の評価実験では重複してクラスタリングされたノードはノード番号 27 の 1 つのみであった．これは，8 カ月間に共通して出現しているブロガが 172 人と少なかったことが原因である．そこで，提案手法のスケーラビリティを評価するために，8 カ月間ではなく，1 月から 3 月までの 3 カ月間のデータを用いて，シミュレーション実験を行った．その結果，3 カ月間に共通して出現するブロガは 751 人，最小支持数を 3 として 1 つの頻出部分グラフを得ることができた．クラスタリングの結果，22 個のクラスタが識別され，147 ノードが複数のクラスタに重複してクラスタリングされた．グラフの規模が大きくなれば，重複してクラスタリングされるノードも増加し，提案手法の有効性が

より発揮されることが分かった．

6. まとめ

本論文では，ブログ空間から頻出かつ重複を許したコミュニティを発見するための手法を提案した．提案手法は，グラフ問題からアイテムセットの問題へ変換し，LCM 法を用いた頻出部分グラフ抽出手法と，Newman 法の応用である，縮約グラフの作成と再クラスタリングによる重複を許したクラスタリング手法である．提案手法の有効性を示すために，収集したデータを用いて評価実験を行った結果，数カ月にわたり，特定の話題に偏ったコミュニティ，つまり頻出コミュニティを抽出できた．また，複数のコミュニティに重複してクラスタリングされるノードの抽出が可能であり，重複を許したクラスタリング手法の妥当性を示すことができた．

今回抽出された頻出コミュニティは，スポーツに関する話題が中心になっていた．これは，記事を収集する際にホスティングサービスを限定しているため，トラックバック数の多い記事を作成しているブロガにたどりつき，局所的な記事の収集に陥ってしまったことが原因であると考えられる．このことから，データ収集法の改善により，より大規模なデータを用いた評価実験を行い，幅広い話題の頻出コミュニティを得ることが今後の課題としてあげられる．

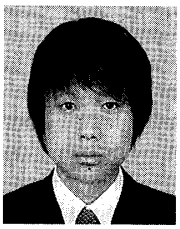
謝辞 本研究の一部は，日本学術振興会・特別研究員奨励費（課題番号：18・00205），日本学術振興会・科学研究費補助金（基盤研究（C）(一般)，課題番号：17500097）の支援により行われた．

参考文献

1) Gruhl, D., Guha, R.V., Liben-Nowell, D. and Tomkins, A.: Information Diffusion Through Blogspace, *WWW*, pp.491–501 (2004).
2) 柴田尚樹，内田 誠：ブログ記事ネットワークにおけるトピックマップの作成，第 2 回ネットワーク生態学シンポジウム論文集 (2006).
3) 谷口智哉，松尾 豊，石塚 満：Blog コミュニティの抽出と分析，第 6 回セマンティックウェブとオントロジー研究会，人工知能学会研究会資料，pp.0801–0806 (2004).
4) Takaki, M., Mori, Y., Tamura, K., Kuroki, S. and Kitakami, H.: Method for Extracting Frequent Communities from Blog User Spaces, *PDPTA*, pp.773–779 (2007).
5) Kuramochi, M. and Karypis, G.: Frequent Subgraph Discovery, *ICDM*, pp.313–320 (2001).
6) Inokuchi, A., Washio, T. and Motoda, H.: An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data, *PKDD*, pp.13–23 (2000).
7) Yan, X. and Han, J.: gSpan: Graph-Based Substructure Pattern Mining, *ICDM*, pp.721–724 (2002).
8) Hu, H., Yan, X., Huang, Y., Han, J. and Zhou, X.J.: Mining Coherent Dense Subgraphs Across Massive Biological Networks for Functional Discovery, *ISMB* (*Supplement of Bioinformatics*), pp.213–221 (2005).
9) Uno, T., Kiyomi, M. and Arimura, H.: LCM ver. 2: Efficient Mining Algorithms for Frequent/Closed/Maximal Itemsets, *FIMI* (2004).
10) Clauset, A., Newman, M.E.J. and Moore, C.: Finding community structure in very large networks, *Physical Review E*, Vol.70, p.066111 (6 pages) (2004).
11) Flake, G.W., Lawrence, S. and Giles, C.L.: Efficient identification of Web communities, *KDD*, pp.150–160 (2000).
12) Flake, G.W., Lawrence, S., Giles, C.L. and Coetzee, F.: Self-Organization of the Web and Identification of Communities, *IEEE Computer*, Vol.35, No.3, pp.66–71 (2002).
13) Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment, *SODA*, pp.668–677 (1998).
14) Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A.: Trawling the Web for Emerging Cyber-Communities, *Computer Networks*, Vol.31, No.11-16, pp.1481–1493 (1999).
15) 豊田正司，喜連川優：日本におけるウェブコミュニティの発展過程，*DBSJ Letters*, Vol.2, No.1, pp.35–38 (2003).
16) Palla, G., Derenyi, I., Farkas, I. and Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society, *Nature*, Vol.435, No.7043, pp.814–818 (2005).
17) Zhang, S., Wang, R.-S. and Zhang, X.-S.: Identification of overlapping community structure in complex networks using fuzzy c-means clustering, *Physica A Statistical Mechanics and its Applications*, Vol.374, pp.483–490 (2007).
18) Dunn, J.C.: A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, *Cybernetics*, Vol.3, pp.32–57 (1973).
19) Bezdek, J.C.: *Pattern Recognition with Fuzzy Objective Function Algorithms*, Kluwer Academic Publishers, Norwell, MA, USA (1981).
20) Reichardt, J. and Bornholdt, S.: Detecting

fuzzy community structures in complex networks with a Potts model, *Physical Review Letters*, Vol.93, p.218701 (2004).
21) Reichardt, J. and Bornholdt, S.: Statistical Mechanics of Community Detection, *Physical Review E*, Vol.74, p.016110 (2006).

(平成 19 年 8 月 3 日受付)
(平成 19 年 9 月 20 日再受付)
(平成 19 年 11 月 11 日採録)



高木　允（学生会員）

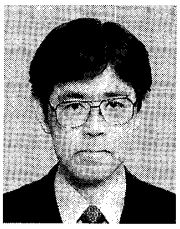
2005 年広島市立大学大学院情報科学研究科知能情報システム工学専攻博士前期課程修了．現在，同大学院情報科学研究科情報科学専攻博士後期課程に在学中．2006 年より日本学術振興会特別研究員 DC．データマイニング，並列分散処理に興味を持つ．日本データベース学会学生会員．

森　康真（正会員）

1994 年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．現在，広島市立大学大学院情報科学研究科助教．データマイニングの研究に従事．電子情報通信学会，人工知能学会，日本データベース学会，IEEE CS，ACM 各会員．

田村　慶一（正会員）

1998 年九州大学工学部情報工学科卒業．2000 年同大学大学院システム情報科学研究科知能システム学専攻修士課程修了．2003 年同大学院システム情報科学府知能システム学専攻博士後期課程単位取得のうえ満期退学．博士（情報科学）．現在，広島市立大学大学院情報科学研究科助教．データベースとその並列分散処理に関する研究に従事．日本データベース学会，IEEE CS 各会員．



北上　始（正会員）

1976 年東北大学大学院工学研究科博士前期課程修了．同年富士通株式会社入社．以後，富士通研究所，新世代コンピュータ技術開発機構，国立遺伝学研究所客員助教授を経て，現在，広島市立大学大学院情報科学研究科教授．データマイニング，生命情報学，知識ベース等の教育研究に従事．博士（工学）．情報処理学会 25 周年記念論文，日本工学教育協会論文論説賞，情報処理学会一般情報処理教育委員会委員，人工知能学会評議員，電子情報通信学会データ工学ワークショップ DEWS2008 組織委員，日本データベース学会 BI 研究グループ運営委員，情報処理学会（TOM）論文誌編集委員，IEEE，ACM，電子情報通信学会各会員．