2007－MPS－66 (13)
2007／9／3

# 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[†] 田村慶一[††]
森 康真[††] 北上 始[††]

本研究では，ブロガーをノード，トラックバックによる繋がりを辺としたグラフから，数ヶ月に渡って頻出し，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，頻出部分グラフを抽出し，頻出部分グラフに重複を許したクラスタリング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，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・関心を持つ頻出なコミュニティと，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ーを発見できた．

# Extraction Method of Overlapping Frequent Communities from Blog User Spaces

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

In this study, we propose a technique which extracts frequent and overlapped communities across multiple months from graphs. 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. After that, LCM algorithm is applied to extracting the frequent itemsets. Finally, we applied the Newman's algorithm to finding overlapping clusters. To confirm the availability of proposed technique, we extracted the frequent communities. As a result, frequent communities and the bloggers who are clustered multiple clusters are extracted.

## 1. はじめに

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上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．

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ーに着目し，ブロガーをノード，記事のトラックバックに基づくブロガー同士の繋がりを辺とみなしたグラフ構造に着目している[8]．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，頻出なコミュニティとは，時系列グラフから抽出される頻出な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，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．

本稿では，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) 時系列グラフからの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

(1) では，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 LCM アルゴリズム[9] 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．

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

(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[1]（以下，Newman 法）を応用し，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，特定の話題で繋がっているだけでなく，コミュニティ内でなんらかの交流関係や社会的な繋がりを持つコミュニティであると考えられる．

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，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ーを発見できた．

論文の構成は以下の通りである．2 章で関連研究について述べる．3 章で提案手法について説明し，4 章で評価実験の結果と考察を示す．5 章でまとめる．

## 2. 関連研究

本章では，「コミュニティの発見・解析」，「グラフマイニング」，「重複を許したクラスタリング」の 3 つの視点から関連研究を述べる．

### 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[2),5)]では，Web ページ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．即ち，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繋がりを持つ，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

グラフデータベース $D = \{G_1, \cdots, G_n\}$ から頻出部分グラフを抽出する研究[4),6)]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．

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グラフであり，一般グラフとは異なる．また，各 $G_i$ のノード数が数百ノード以上の規模であるため，文献 4), 6) の手法を，我々が想定している問題に応用するには計算時間の面で不向きである．

文献 3) で提案されている CODENSE を，著者らが収集したデータに応用した実験では，本来ひとつになるべきコミュニティがサイズの小さな部分グラフに分割されたという問題が生じている．

そのため，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，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，ネットワーク構造解析に基づくクラスタリング手法が盛んに行われている[7),10)]．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頂点に置き換えたものであるといえ，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繋がり方を把握できるという利点を持っている．

Palla ら[7)] や Zhang ら[10)] はネットワーク構造に基づくクラスタリング手法に，重複を許したクラスタリングを取り入れている．しかしながら，両者において，閾値の設定が必要であり，最適な閾値を決定する指導原理的な指標は存在しない．

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. 提案手法

ブロガーをノード，トラックバックによる繋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，$u, v$ はひとつのノード，$\{u, v\}$ 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 1 (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 つのステップから成る．

(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得る重複を許したクラスタの抽出

### 3.1 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問題を頻出アイテム集合抽出問題に変換し，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．例を用いた説明を図 1 に示す．

前処理として，$D = \{G_1, \cdots, G_n\}$ から全ての $G_i$ に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 1 (a) 中の塗りつぶされているノードを全ての $G_i$ に共通しているノードとすると，図 1 (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 1 (b) のグラフデータベース $D'$ が得られる．

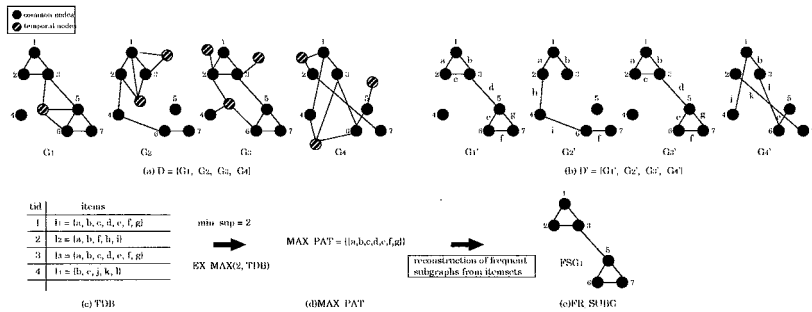


図 1　頻出部分グラフ抽出例

次に，問題変換のために，グラフの辺にラベルを付与する（図 1 (b)）．各 $G'_i$ について，ラベル集合 $I_i$ を作成する．このラベル集合をアイテム集合とみなし，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 1 (c)）．

作成した $TDB$ から，LCM 法を用いて極大頻出アイテム集合を抽出する．ここで，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．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は，$MAX\_PAT = \{PAT_1, \cdots, PAT_l\}$ となる（図 1 (d)）．$PAT_i$ を用いて，頻出アイテム集合から元のグラフの復元を行うと，図 1 (e) に示すような頻出部分グラフが得られる．

### 3.2　重複を許したクラスタリング手法

提案するクラスタリングアルゴリズムは，Newman 法でクラスタリング後，縮約グラフを作成し，再度 Newman 法でクラスタリングを行うことで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，Newman 法を関数 $Newman(G)$ と表記する．また，図 2 に重複を許したクラスタリングの例を示す．

まず，図 2 (a) に示すように，$Newman(G)$ を用い，入力グラフ $G$ 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それぞれ $NC_i$ とする．それぞれの $NC_i$ について，関数 $MAKE\_C\_GRAPH(G, NC_i)$ を用いて，$NC_i$ をひとつのノードとした縮約グラフを作成する（図 2 (b)）．このとき，ひとつのノードに縮約されたクラスタに繋がっている辺の数の情報はそのまま保持しておく．$MAKE\_C\_GRAPH(G, NC_i)$ の出力として，縮約グラフ $SG_i$ が得られる．このとき，縮約されたクラスタを $v_i$ とする．

次に，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発見する．再クラスタリングする際には，Newman 法で使用しているクラスタリングの評価指標である $Q$ の値が増加するノードのみをクラスタリングする．その結果，重複してクラスタリングされるノード集合 $DUP_i$ が得られる（図 2 (c)）．得られたノード集合を $NC_i$ に付加する．この手順を全ての $NC_i$ について行い，重複してクラスタリングされる全てのノードを見つけ出す．

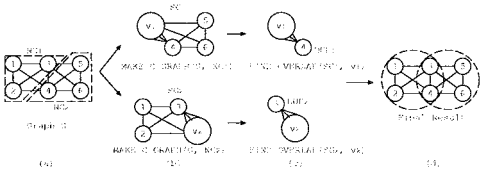

図 2　重複を許したクラスタリングの例

## 4. 評価実験

評価実験では，文献 8) で用いたデータ収集法を使用した．また，収集したデータは 2006 年 1 月から 2006 年 8 月までの 8ヶ月間のデータを使用した．つまり，時系列グラフデータは，$D = \{G_1, \cdots, G_8\}$ となる．

$D$ に前処理を施すと，全ての $G_i$ に共通して出現しているノードは 172 ノードであった．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．

時系列グラフ $D$ に提案手法を適用したところ，1 つ頻出部分グラフが抽出され，重複を許したクラスタリングを行うことにより，11 個のクラスタが識別された．ここで，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ーが書いたブログ記事に対して，$tf - idf$ 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 3 件を表 1 に示す．

表 1 から，各クラスタとも野球，サッカーの話題に

表1 各クラスタにおける *tf-idf* による解析結果

| クラスタ＼ランク | 1位 | 2位 | 3位 |
|---|---|---|---|
| 1 | ホークス | 野球 | ブログ |
| 2 | マリーンズ | トラック | ブログ |
| 3 | カープ | 広島 | 町 |
| 4 | ライオンズ | 野球 | イタリア |
| 5 | ライオンズ | ステージ | ブログ |
| 6 | 野球 | ライオンズ | ブログ |
| 7 | ブログ | スワローズ | 東京 |
| 8 | 日本 | ドイツ | サッカ |
| 9 | ブログ | 阪神 | 日記 |
| 10 | 日本 | ブログ | ドイツ |
| 11 | 中日 | ドラゴンズ | 日本 |

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ーの記事を調べてみたところ，$tf-idf$ 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ーが各クラスタ内に多数存在していた．以上より，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言うことが可能である．

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．本実験では，ノード 27 がクラスタ 3 とクラスタ 10 に重複して所属する結果となった．図 3 にクラスタリング結果を示す．

ここで，ノード 27 のブロガー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 のブロガーが保持している記事から，$tf-idf$ を用いて重要なキーワードの抽出を行った．結果として，1．広島，2．カープ，3．ブログというキーワードが得られた．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．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．ノード 27 のブロガー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や，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．

## 5. 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．評価実験を行った結果，数か月に渡り，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．

今後の課題として，より大規模なデータを用いた評価実験が挙げられる．

**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・0205），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．

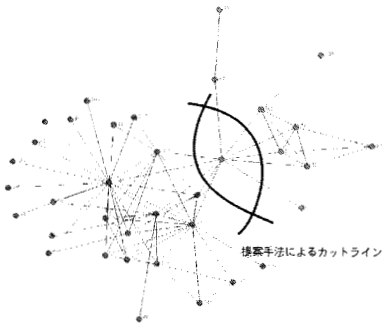


図 3 提案手法によるクラスタリング結果

## 参 考 文 献

1) A.Clauset, M.E.J. Newman, and C.Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111(6 pages), 2004.
2) D. Gruhl, R. V. Guha, D. Liben-Nowell, and A. Tomkins. Information Diffusion Through Blogspace. In *WWW*, pages 491–501, 2004.
3) H. Hu, X. Yan, Y. Huang, J. Han, and X. J. Zhou. Mining Coherent Dense Subgraphs Across Massive Biological Networks for Functional Discovery. In *ISMB*, pages 213–221, 2005.
4) A.Inokuchi, T.Washio, and H.Motoda. An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data. In *PKDD*, pages 13–23, 2000.
5) J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *SODA*, pages 668–677, 1998.
6) M. Kuramochi and G. Karypis. Frequent Subgraph Discovery. In *ICDM*, pages 313–320, 2001.
7) G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.
8) M.Takaki, Y.Mori, K.Tamura, S.Kuroki, and H.Kitakami. Method for extracting frequent communities from blog user spaces. In *PDPTA*, pages 773–779, 2007.
9) T. Uno, M. Kiyomi, and H. Arimura. LCM ver. 2: Efficient Mining Algorithms for Frequent/Closed/Maximal Itemsets. In *FIMI*, 2004.
10) S. Zhang, R.-S. Wang, and X.-S. Zhang. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A Statistical Mechanics and its Applications*, 374:483–490, Jan. 2007.