2D-3

ブログユーザ空間からの頻出な部分グラフの抽出

高木 允† 森 康真‡ 田村 慶一‡ 黒木 進‡ 北上 始‡
広島市立大学大学院†/日本学術振興会 広島市立大学‡

1. はじめに

近年，ブログ上からの知識発見に関する研究が様々行われている．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みなした複数のグラフから頻出な部分グラフを抽出し，Newman らによって提案されている手法[1]によりクラスタリングする手法を提案する．

提案する手法により，長期間に渡り形成されている，強い繋がりを持ったコミュニティを発見できる．コミュニティ内の話題に興味を持つブロガーへの情報推薦などの応用が期待できる．実際にデータを収集し，提案手法を適用した結果，長期間に渡って形成されているコミュニティを発見できた．

2. データ収集

図 1 に示すように，始点ノードである記事をランダムに選択し，トラックバックを辿ることにより記事を収集していく．記事の URL からブロガー名を特定し，重みなし，無向グラフを作成する．同一ブロガーが再度トラックバックされている場合は新たなノードを作成せず，辺のみを付け加えていく．最終的に生成されるグラフのノード数は収集したブロガー数と等しくなる．記事の収集を，「2006 年 6 月 1 日から 2006 年 6 月 30 日までに投稿された記事」のように 1 ヶ月単位で行った．データ収集を 2006 年 6 月から 2006 年 9 月までの 4 ヶ月間行った．つまり，4 つのブロガーのグラフが生成されることになる．

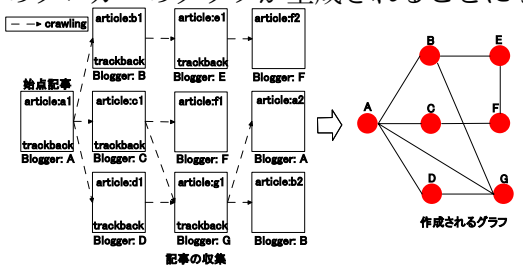

図 1 データ収集とグラフ作成

3. 記号定義

収集した複数のグラフからグラフデータベース $D=\{G_1,\ldots,G_n\}$ を作成する．G_i はグラフを表現しており，$G_i=G(V_i,E_i)$ と定義する．V_i はノードの集合，E_i はノードのペアにより辺を表した辺の集合である．頻出部分グラフを抽出するために，D から全ての G_i に共通しているノードを抽出したグラフ $D'=\{G_1',\cdots,G_n'\}$ を作成する．

Extraction of Frequent Subgraphs from Blog User Space
†Makoto TAKAKI, Graduate School of Hiroshima City University / JSPS
‡Yasuma Mori, Hiroshima City University
‡Keiichi TAMURA, Hiroshima City University
‡Susumu KUROKI, Hiroshima City University
‡Hajime KITAKAMI, Hiroshima City University

$G_i'=G(V',E_i')$ であり，$V'=V_1\cap\ldots\cap V_n$，E_i' は V' に含まれるノードのみで構成された辺の集合である．E_i' の全ての和を E' とする．$|E'|$ 個のラベルを要素とした I を作成し，E' から I への全単射を f とすると関数 f は以下のように定義できる．

$$f:E'\rightarrow I \quad \text{または} \quad I=f(E') \qquad (1)$$

式（1）を用いて各 E_i' にラベル付けを行い，ラベルをアイテムとしたアイテム集合を $I_i=\{label_{i1},\ldots,label_{im}\}$ とする．トランザクションデータベースを $TDB=\{t_1,\ldots,t_n\}$ と定義する．ここで $t_i=(i,I_i)$ である．

最終的に抽出された頻出部分グラフを $FSG_i=G(FV_i,FE_i)$ とする．ここで，FE_i は頻出な辺の集合であり，FV_i は FE_i を構成する全てのノードの集合である．

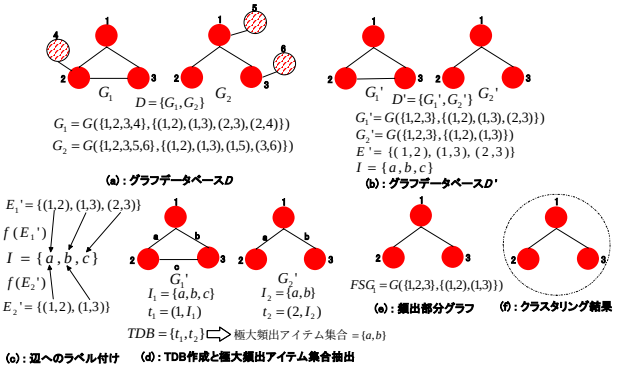

図 2 提案手法の概要

4. 提案手法

本研究では，複数のブロガーのグラフから頻出な部分グラフを抽出し，クラスタリングを行うことでコミュニティを発見する．図 2 に提案手法の概要を示す．以下，提案手法の処理手順を示す．

（1）図 2（b）に示すように，図 2（a）の G_1，G_2 からノード 1，2，3 を抽出し，D' を作成する．全ての E_i' の和集合 E' を作成し，$|E'|$ 個の要素を持ったラベル集合 I を作成する．図 2（b）では $E'=\{(1,2),(1,3),(2,3)\}$，$I=\{a,b,c\}$ である．関数 f を定義し，各辺とラベルを対応付ける．

（2）関数 f を用いて各 E_i' の辺とラベルを対応付けて（図 2（c）），ラベルをアイテムとしたアイテム集合 I_i を作成する．図 2（d）においては，$I_1=\{a,b,c\}$，$I_2=\{a,b\}$ となる．

（3）I_i を用いてトランザクションデータベース TDB を作成する．作成した TDB から，文献[2]で提案されている手法を用いて極大頻出アイテム集合を抽出する．図 2 では，極大頻出アイテム集合として{a,b}が得られる．得られたアイテム集合から f^{-1} を用いて辺を復元する．復元された辺からノード集合を復元し，頻出部分グラフ FSG_i を得る（図 2（e））．

（4）図 2（f）に示すように，復元された FSG_i を Newman らによって提案されているクラスタリング手法を用いてクラスタリングする．Newman らのアルゴリズムは，ノード集合を辺の繋がりにより分割していくクラスタリング手法である．FSG_i をクラスタリングし，コミュニティを発見する．

本手法の特長は，ブロガーのグラフに直接 Newman らのアルゴリズムを適用するのではなく，頻出な部分グラフを取り出して Newman らのアルゴリズムを適用することで，より繋がりの強いブロガー集団を見つけ出せることである．

5. 評価実験

実際にデータを収集し，グラフデータベース $D=\{G_1,G_2,G_3,G_4\}$を作成した．D から，4 ヶ月に渡って共通して出現しているブロガーを抽出し，グラフデータベース $D'=\{G_1', G_2', G_3', G_4'\}$を作成する．すべての月に存在していたブロガーの数は 319 人であった．つまり，$|V'|=319$ であり，$|E_1'|=1{,}650$ ， $|E_2'|=1{,}695$ ， $|E_3'|=1{,}697$, $|E_4'|=1{,}230$ であった．

5.1. D のクラスタリング結果

D 中の 6 月のブロガーのグラフ G_1 に Newman らが提案しているアルゴリズムを適用した結果，43 個のクラスタが識別された．クラスタサイズは最大で 1195，最小で 2 であり，サイズが 10 未満のクラスタが 29 個，サイズが 400 以上のクラスタが 6 個，残りのクラスタはサイズが 16～156 であった．極端に小さなクラスタが多数存在し，極端に大きなクラスタと中間サイズのクラスタは少数であった．

各クラスタについて *tf-idf* を用いた解析を行った結果，ある野球チームの話題を主としているブロガー，政治の話題を主としているブロガーのように，様々なブロガーが混在していた．トラックバックを調査すると，特定のイベントのために発生している一過性のトラックバックが多数存在した．一過性のトラックバックが多く混在しており，共通の興味・趣味を持ったブロガー集団のコミュニティの発見が困難となることが分かる．

5.2. D' のクラスタリング結果

D' に提案した手法を適用した．頻出部分グラフを抽出するための最小支持数は 2 とし，抽出された頻出部分グラフ FSG は全部で 6 個あった．ここでは，抽出された FSG_1 について説明する．FSG_1 をクラスタリングした結果，13 個のクラスタが識別された．表 1 に各クラスタの記事に *tf-idf* を適用し，*tf-idf* の値の上位 3 件のキーワードを示す．表中の $CLUSTER_{ij}$ はクラスタリングされた個々のクラスタの識別子を表している．ここでは，4 つのクラスタの *tf-idf* 上位 3 件を示している．図 3 に結果を可視化したものを示す．

表 1 から，*tf-idf* によって抽出されたキーワード上位 3 件はそれぞれ容易に連想できるキーワードとなっている（例えば，表 1 の $CLUSTER_{11}$ はプロ野球のカープについての記事を扱っている集団である）．手作業でクラスタに属しているブロガーのブログを確認したところ，$CLUSTER_{113}$ では，阪神ファンのブロガーが 90% を占めていた．ファンの判断基準としては，ブログの題名やプロフィールなどから，自ら阪神ファンであることを記述しているブロガーをファンであると判断した．このようにクラスタリング結果とクラスタを解析した結果が強い相関を持っているのは，頻出な部分グラフを抽出することで一過性のトラックバックによるブロガー間の繋がりを除去することができ，より繋がりの強いブロガー同士の繋がりのみを抽出し，クラスタリングできたためである．

表 1　各クラスタの *tf-idf* 上位 3 件

	tf-idf 上位 3 件		
クラスタ	1	2	3
$CLUSTER_{11}$	カープ	広島	日本
$CLUSTER_{12}$	投手	楽天	野球
$CLUSTER_{13}$	楽天	イーグルス	野球
$CLUSTER_{14}$	日本	ブラジル	ドイツ

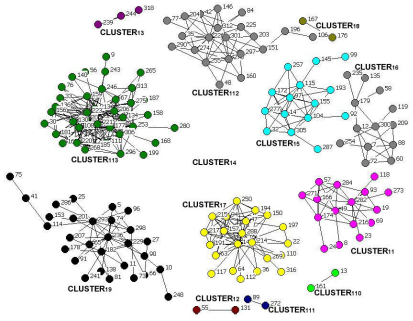

図 3　FSG_1 のクラスタリング結果

6. まとめ

本論文では，辺へのラベル付けを行って頻出部分グラフを抽出し，個々の頻出部分グラフをクラスタリングする手法を提案した．実際に収集したブログデータに提案手法を適用した．収集したデータそのものをクラスタリングした場合と比較すると，提案手法では，より精度の高いクラスタリングが可能であることが分かった．さらに，複数ヶ月に渡って同一の興味・関心を持っているブロガー集団を発見できた．

参考文献

[1] M. E. J. Newman. Fast Algorithm for Detecting Community Structure in Neworks. Physical Review E, Vol. 69, p.066133, 2004.

[2] Takeaki Uno, Masashi Kiyomi, and Hiroki Arimura. LCM ver.2: Efficient Mining Algorithms for Frequent/Closed/Maximal Itemsets. In *FIMI*, 2004.