大規模グラフ系列として捉えたソーシャルブックマークデータからのコミュニティ変化ルール抽出

山口　雄大† 　新美　礼彦†† 　小西　修†††

† 公立はこだて未来大学大学院 システム情報科学研究科
†† 公立はこだて未来大学 システム情報科学部 情報アーキテクチャ学科
††† 公立はこだて未来大学 システム情報科学部 複雑系知能学科
〒 041–8655 北海道函館市亀田中野町 116 番地 2
E-mail: †{g2109046,niimi,okonishi}@fun.ac.jp

あらまし　本研究では，データストリームのトランザクションデータ集合を，ある期間ごとのデータ関係構造とその変化を表すグラフ系列と捉え，そのグラフ系列中のコミュニティの変化を解析する．本研究が提案するアルゴリズムでは，拡張グラフカーネルと階層的クラスタリングを組み合わせた手法によって，グラフ系列全体でコミュニティの関係を解析し，不定期に出現する (系列の途中で見えなくなる) コミュニティの変化をルールとして抽出する．そして，人工データセットと実際のソーシャルブックマークデータを用いた評価実験の結果から，提案アルゴリズムが不定期に出現するコミュニティの変化を捉えられることが示された．

キーワード　グラフ系列, コミュニティ変化ルール，グラフカーネル, クラスタリング, ソーシャルブックマーク

Extraction of Community Transition Rules from Social Bookmark Data as Large Graph Sequence

Takehiro YAMAGUCHI†, Ayahiko NIIMI††, and Osamu KONISHI†††

† Graduate School of Systems Information Science, Future University Hakodate
†† Department of Media Architecture, Faculty of Systems Information Science, Future University Hakodate
††† Department of Complex and Intelligent Systems, Faculty of Systems Information Science, Future University Hakodate
116–2 Kamedanakano, Hakodate, Hokkaido, 041–8655 Japan
E-mail: †{g2109046,niimi,okonishi}@fun.ac.jp

Abstract　In this study, we treat transactional sets of a data stream as a graph sequence. This graph sequence represents both the relational structure of data for each period and changes in these structures. In addition, we analyze changes in a community in this graph sequence. Our proposed algorithm extracts community transition rules, in order to detect the communities that appear irregularly in a graph sequence, using our proposed method combined with an extended graph kernel and hierarchical clustering. In the experiment using synthetic datasets and social bookmark datasets, we demonstrated that our proposed algorithm could detect changes in a community appearing irregularly.

Key words　Graph Sequence, Community Transition Rules, Graph Kernels, Clustering, Social Bookmark

1. はじめに

近年，新しいタイプの大規模データとしてデータストリームが注目されている．データストリームとは,「膨大な量のデータが，高速なストリームを通じて，時間的に変化しながら，終わりなく到着し続ける」という特性を持つ動的な大規模データである [10]．現在，これらの蓄積されたデータをいかに分析し，有効活用できる知識を発見するかが重要な課題となっている．

本研究が分析の対象とするデータストリームは，トランザクションデータの集合である．例えば，実際のデータストリーム

としてソーシャルブックマーク (以降，SBM) データを考えた場合，そのデータは図 1 のような一連の属性 (ユーザ ID，ブックマークされた Web ページの URL，ブックマークされた日付，ブックマークに付与されたタグ情報) に対する値を組み合わせたトランザクションデータの集合である．

user_id	url	timestamp	tags
1	http://clip.livedoor.com/	2006-06-27 17:24:54	sbm clip これはすごい web2.0
2	http://clip.livedoor.com/	2006-06-27 17:27:46	sbm
3	http://clip.livedoor.com/	2006-06-27 17:46:44	sbm livedoor
4	http://clip.livedoor.com/	2006-06-27 19:31:15	R18
5	http://clip.livedoor.com/	2006-06-27 20:33:55	livedoor,コンテンツ

図 1　ソーシャルブックマークデータのトランザクション集合

そして，このようなデータストリームのトランザクションデータ集合を，ある期間ごとのデータ関係構造とその変化を表すグラフ系列と捉えて，その特徴的な変化を解析する．再び SBM を例にすると，SBM を利用するユーザをノード，特定の期間内で同じ Web ページをブックマークしているユーザ間の関係をリンクとしたグラフ構造を定義し，ある期間のデータごとにグラフを作成することで，同じページをブックマークしているユーザ関係の変化をグラフ系列として表現することができる (図 2 参照)．つまり，時間の経過と共に変化するユーザ間の関係を，グラフ上のノードやリンクの追加，削除によって表現することができる．

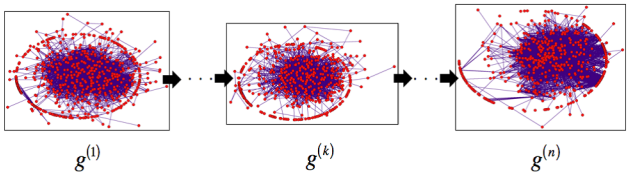

図 2　ソーシャルブックマークのグラフ系列

本研究が解析する特徴的な変化とはグラフ系列中の各グラフに存在する密な部分グラフ構造 (以降，コミュニティ) の変化である．グラフ構造に存在するコミュニティを分析することは，これまでに社会ネットワークや，論文の引用ネットワーク，タンパク質の相互作用ネットワークなど広範囲で行われている．特に近年では，動的なネットワークからコミュニティを抽出し，その進化を解析することが重要な課題となっている．本研究では，データストリームの変化を表すグラフ系列中に存在するコミュニティとその変化を解析することで，より多くのデータと関係を持つデータ構造とそれらの変遷を発見できると考える．

しかし，このグラフ系列中の各グラフは，特定の期間に存在するデータのみを用いて形成される関係構造であるため，その期間に到着しないデータを表すノードと，そのノードと繋がるリンクが見えなくなってしまう．そのため，系列中の全てのグラフには存在しないが，ある特定の期間だけ存在するコミュニティが多数出現することが考えられる．そこで本研究では，そのようなグラフ系列中で不定期に出現するコミュニティの変化を捉えるアルゴリズムを提案する．我々の知る限り，「グラフ系列中の連続しないグラフ間で現れる同じ構造を持ったコミュニティ」を解析対象とした従来研究はない．そして，コミュニティが生成されてから消滅するまでの変化の過程を変化ルールとして抽出することで，それらのコミュニティで盛り上がった話題の変化を解釈できると考える．

2. 関 連 研 究

近年，様々なシステムをネットワークとして表現し，そのネットワークに存在する高密度なリンク構造の特徴を抽出しようとする研究が多く行われている．この高密度なリンク集合は「コミュニティ」とも言われ，これまでに社会ネットワークや，論文の引用ネットワーク，タンパク質の相互作用ネットワークなどの様々な分野で，その構造の分析や計算量を抑えた抽出手法が議論されてきた．

特に現在では，動的なネットワークからコミュニティを抽出し，その進化を研究することが重要な課題となっている．コミュニティの進化や時間的な変化を解析する先行研究として，ウェブコミュニティの全体的な発展過程を分析する研究 [12] や，異なるノードタイプを含む書誌情報ネットワークの進化を分析する研究 [6]，コミュニティの抽出とその進化を分析するための体系的なフレームワークを提案した研究 [9] などが挙げられる．また，ネットワークからのコミュニティ抽出は，グラフを最適な部分グラフに分割する，クラスタリングの問題と捉えることもできる．そして，動的なグラフ構造のクラスタリングに対して，時間的な滑らかさを組み込んだスペクトルクラスタリングのアルゴリズムが提案されており，より安定的で一貫性のあるクラスタリング結果を得られることが示されている [8]．

これらの関連研究では，時刻 t のネットワークからコミュニティ構造を抽出する場合,「時刻 $t-1$ のコミュニティ構造から劇的に逸脱しない」かつ「時刻 t のデータにより適した」コミュニティ構造を，最適な抽出結果と評価している．しかし，本研究が定義したデータストリームの変化を表すグラフ系列は，特定の期間に到着するデータのみを用いて形成されるグラフから成るグラフ系列であるため，その期間に到着しないデータを表すノードとそのノードと繋がるリンクが見えなくなってしまう．そのため，系列中の全てのグラフには存在しないが，ある特定の期間だけ存在するコミュニティが多数出現することが考えられる．つまり，それらのコミュニティは見かけ上，時刻 t において「時刻 $t-1$ のコミュニティ構造から劇的に逸脱する」コミュニティ構造となる．したがって，関連研究の手法ではそのようなコミュニティの変化を捉えることが困難であると考える．そこで本稿では，グラフ系列全体でコミュニティの関係を解析するアルゴリズムを提案し，不定期に出現するようなコミュニティの変化を抽出できることを示す．

3. 提 案 手 法

本研究が提案するアルゴリズムでは，データストリームのトランザクション集合を入力データとし，以下の 5 つのステップに従って，コミュニティの変化ルールを出力する．1) 入力されたデータストリームのトランザクション集合からグラフ系列を作成し，2) 系列中のグラフ毎にコミュニティを抽出する．次に，3) 拡張グラフカーネルを用いて，コミュニティ間の類似度を算出し，4) その類似度を基にコミュニティクラスタを作成する．そして，5) コミュニティクラスタから，コミュニティの変化ルールを出力する．本章では，これらの各ステップを詳細に

述べる．

3.1 データストリームのグラフ系列作成

本論文で定義するグラフ系列中の各グラフは，過去のデータ間の関係を保持せず，特定の期間内に限定されたデータの関係構造を表す．そのため，各グラフに含まれるデータの期間は抽出されるコミュニティの変化ルールに大きく影響を与えると考えられる．従って，実際のデータストリームのトランザクションデータ系列からグラフ系列を作成する場合，解析対象のデータストリームごとの性質に特化したグラフ系列手法が求められる．本節では，評価実験で用いる SBM データに特化したグラフ系列作成手法を示す．

本論文では，SBM を日常的に利用しているユーザのブックマーク行動を系列中の全てのグラフ上で表現できるように，バーストを考慮したグラフ系列の作成を行う．本論文で定義するバーストとは，ある瞬間に爆発的に利用されて始めた話題によって，トランザクション数が急激に上昇する現象のことである．そのような話題が発生した場合，日常的に SBM を利用するユーザのブックマーク行動が活発となり，トランザクション数が他の日に比べて，上昇することが予想される．そのトランザクション数の違いを考慮してグラフ系列を作成する．グラフ系列の具体的な作成手順を以下に示す．

(1) データストリームのトランザクション数を日別に観測する．

(2) 設定した一日のトランザクション数の閾値 BT 以上のトランザクション数が観測された日，観測開始日，観測終了日を，グラフ作成ポイントとする．

(3) 隣接するポイント間のデータごとにグラフを作成し，最も過去のデータ構造を表すグラフから順番にグラフ番号を割り当て，それらをグラフ系列とする．

例えば，日付ごとに整理したデータに対して，$BT = 5000$ と設定した例を図 3 に示す．この場合，BT 以上のトランザクション数を観測した日が 3 日 (横軸：2，5，8) あり，グラフ作成ポイントは観測開始日，観測終了日と合わせて 5 つとなる．そして，隣接するポイント間に含まれるデータごとにグラフを作成し (図の上部：①，②，③，④)，4 つのグラフからなるグラフ系列が作成される (全てのトランザクションデータがグラフ系列上で表現される)．このような観点でグラフ系列を作成することによって，系列中の各グラフにある瞬間に盛り上がった話題に関するユーザ関係構造を表現できると考える．

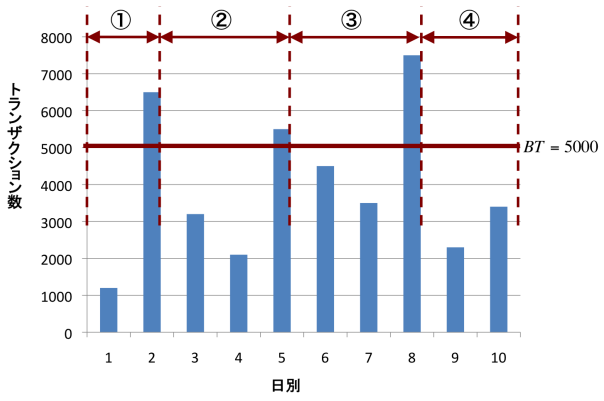

図 3 閾値 BT の適用例

他のデータストリームに対してもこのように，解析対象のデータに沿った観点でグラフ系列を作成することで，データストリームのトランザクションデータ系列に潜在する主要なデータ構造を同じグラフ上に表現できると考える．また，この閾値 BT はコミュニティの変化ルールを出力するのに必要な計算量に影響を与えると考える．例えば，閾値 BT に大きな値を設定すると作成されるグラフ系列が短くなり，必要な計算量が少なくなると考える．

3.2 コミュニティ抽出

本研究が定義するコミュニティとは，モジュラリティの最大化を目指すアルゴリズムによって抽出される，リンク密度の高い部分グラフのことである．モジュラリティとは，分割されたネットワークの評価指標であり，対象とするネットワーク全体をどれだけバランスよくリンク高密度集団に分割したのかを評価している [5]．例えば，対象とするグラフが $C_1, C_2, \ldots, C_L$ と L 個の重複しないコミュニティに分割されたときに，モジュラリティ Q は以下のように定義される．

$$Q = \sum_{l \in \{1,\ldots,L\}} Q_l = \sum_{l \in \{1,\ldots,L\}} (e_{ll} - a_l^2) \qquad (1)$$

e_{ll} は C_l 内部のリンクの存在確率を意味し，a_l は，無向グラフであっても敢えて出・入リンクとして「リンク端」を分けて考えたとき，C_l 内にあるリンク端総数のグラフ全体に対する存在確率を意味する．グラフ中のリンクの総数を m としたとき，e_{ll} と a_l はそれぞれ以下の式で得られる．

$$e_{ll} = \frac{1}{2m} \sum_{i \in V_l} \sum_{j \in V_l} A(i,j) \qquad (2)$$

$$a_l = \frac{1}{2m} \sum_{i \in V_l} \sum_{j \in V} A(i,j) \qquad (3)$$

ここで，V_l はコミュニティ C_l に含まれるノード集合を，V はネットワーク全体に含まれるノード集合を表す．そして，$A(i,j)$ はグラフの隣接行列で，ノード i,j 間にリンクがあると 1，なければ 0 を返し，グラフ中のリンク端の総計は $2m$ である．つまり，e_{ll} は各コミュニティ内部の密度がそれぞれ高いことを求め，a_l は全体をひとつのコミュニティにしている場合や，ランダムな分割に対して Q を下げる補正項として導入されている．

我々は，モジュラリティを最大化する手法のひとつである *CNM(Clauset-Newman-Moore)* 法 [1] を用いて，グラフ系列中のグラフ毎にコミュニティを抽出する．この手法は，グラフ中の各ノードを 1 ずつの仮コミュニティとみなして，Q が最も増える 2 つの仮コミュニティを統合し，それを Q が最大になるまで繰り返す．また，この手法は圧倒的に計算量のオーダーが小さく，パラメータ調整が不要である特性を持っており，それがこの手法を用いる根拠となる．そして，抽出されたコミュニティに対してユニーク ID，C_i^k を割り当てる．これは，グラフ系列中の k 番目のグラフにおける，あるコミュニティの ID が i であることを表す．

3.3 コミュニティ間の類似度算出

本研究では，コミュニティ間の類似度を算出する手法として，

カーネル法に着目した．カーネル法は，データの非線形構造をとらえる強力な手法として注目されており，分類や識別などに適用され，高い精度を得ている [3]．また，このカーネル法は，2 つの対象のある種の類似度を定義していると考えることができ，これまでに配列やグラフなどの構造を持ったデータを識別するカーネル関数や，データストリームの時間的な変化を特徴として捉えるストリームカーネル [11], [13] が提案されている．本研究では，コミュニティが表すグラフ構造を入力とし，比較するコミュニティの時間的な変化を考慮したカーネル関数を設計することで，時間間隔の離れた (連続しないグラフ間の) コミュニティの関係を発見できると考える．

ここで，コミュニティの類似度を表すカーネル関数を定義する（以降，拡張グラフカーネルとする）．各コミュニティはノード v にラベルのついた無向グラフ $C_i^m = (V, E)$ である (コミュニティ C_i^m はノード集合 V とリンクの集合 E からなり，各ノード $v \in V$ にはラベル $\sigma(v) \in \Sigma$ が振られている)．そのため，グラフ同士の畳み込みカーネル [2] を適用することができる．つまり，部分構造 $S(C_i^m)$ をコミュニティ C_i^m 中のノードパスの集合とした，コミュニティ同士の畳み込みカーネルは以下のように定義できる．

$$K(C_i^m, C_j^n) = \sum_{s \in S(C_i^m)} \sum_{s' \in S(C_j^n)} f(s|C_i^m) f(s'|C_j^n) K_s(s, s') \tag{4}$$

また，$K_s(s, s')$ は，2 つのノードパス $s = (v_1, v_2, \ldots, v_k) \in S(C_i^m)$ と $s' = (v'_1, v'_2, \ldots, v'_l) \in S(C_j^n)$ の間のカーネル関数であり，以下のようにラベルごとのカーネル関数 K_Σ の積で定義される．

$$K_S(s, s') = \begin{cases} \prod_{i=1}^{n} K_\Sigma(\sigma(s_i), \sigma(s'_i)) & (|s| = |s'| \text{ のとき}) \\ 0 & (|s| \neq |s'| \text{ のとき}) \end{cases} \tag{5}$$

さらに，ラベルごとのカーネル関数は，以下のように定義される．

$$K_\Sigma(\sigma, \sigma') = \begin{cases} 1 & (\sigma = \sigma' \text{ のとき}) \\ 0 & (\sigma \neq \sigma' \text{ のとき}) \end{cases} \tag{6}$$

本研究では，カーネル関数を効率良く計算するため，コミュニティ C_i^m 中のノードパスの集合を，各ノードだけ構成される長さ 1 のパスと，各ノードと隣接するノードから構成される長さ 2 のパスに限定した．これは我々の事前研究 [7], [14] において，コミュニティ内で最も次数の高いノードとそれと隣接するノードを探索するだけで，コミュニティ内の約 7 割のノードを網羅したことがその根拠となっている．つまり，最大で長さ 2 のノードパス集合間の部分構造に限定した比較でも，各コミュニティの特徴を捉えた類似度を表現できると考える．

また，式 (4) の $f(s|C_i^m) f(s'|C_j^n)$ はノードパス s，s' に与えられる重みであり，それぞれの部分構造が類似度全体にどれほどの影響力を持つかを表す．本研究では，グラフ系列全体のコミュニティ間の類似度を算出するため，比較するコミュニティの時間間隔を考慮した重みを以下のように定義した．

$$K(C_i^m, C_j^n) = \sum_{s \in S(C_i^m)} \sum_{s' \in S(C_j^n)} |m - n| \times \lambda^{|s|} \lambda^{|s'|} K_s(s, s') \tag{7}$$

上式は部分構造の重みを，比較するコミュニティのグラフ間隔とノードパス s の長さによって減衰する重みの積と表している．つまり，より長い期間変化しない部分構造が類似度全体に与える影響力を高くしている．そして，グラフ系列の規模に応じて，コミュニティの変化を解析する閾値 CT を定義し，比較するコミュニティのグラフ間隔がこの閾値に含まれない ($|m - n| > CT$) とき，コミュニティ間のカーネル値を 0 とする．さらに，コミュニティ C_i^m と C_j^n の規模の違いを考慮して，以下の式で正規化を行い，これを拡張グラフカーネルの最終出力値とする．

$$K'(C_i^m, C_j^n) = \frac{K(C_i^m, C_j^n)}{\sqrt{K(C_i^m, C_i^m) K(C_j^n, C_j^n)}} \tag{8}$$

3.4 コミュニティクラスタの作成

拡張グラフカーネルによって算出されたコミュニティ間の類似度を基に，コミュニティクラスタを形成する．ここで言うコミュニティクラスタとは，類似度の高いコミュニティ集合のことを指す．本研究では，グラフ番号に関係なく，類似度の高いコミュニティ群を発見するために，ボトムアップなクラスタリング手法 (群平均法) を用いる．そして，クラスタリングの距離関数 $D(Cl_1, Cl_2)$ を以下のように定義する．

$$D(Cl_1, Cl_2) = \frac{1}{n_1 n_2} \sum_{C_i^m \in Cl_1} \sum_{C_j^n \in Cl_2} K(C_i^m, C_j^n) \tag{9}$$

ここで，n_1，n_2 はそれぞれ，クラスタ Cl_1，Cl_2 に含まれるコミュニティの数を表し，$K(C_i^m, C_j^n)$ は系列中の m 番目のグラフにおけるコミュニティ i と，系列中の n 番目のグラフにおけるコミュニティ j との類似度を表す．また，クラスタ併合をクラスタ数が 1 になるまで繰り返すのでなく，同じグラフ番号を持つコミュニティが同一のコミュニティクラスタに含まれることになる，1 つ前の時点でクラスタ併合を終了する．つまり，形成される各クラスタ内で，同じグラフ番号を持つコミュニティが含まれないようにコミュニティクラスタが形成される．

3.5 変化ルールの抽出

作成された全てのコミュニティクラスタに対して，同じクラスタに含まれるコミュニティ群を同一のコミュニティが変化したコミュニティ群と解釈し，そこから変化ルールを抽出する．具体的には，クラスタ内に含まれる各コミュニティをグラフ系列番号を基に時系列に並べることで，以下に定義する「コミュニティの変化」と照合することでき，コミュニティの変化ルールが抽出される．

（1） 生成: コミュニティクラスタにグラフ番号 k のコミュニティが含まれていて，グラフ番号が k より小さいコミュニティが含まれていない．

（2） 拡大: コミュニティクラスタにグラフ番号 k のコミュニティとグラフ番号 $k-1$ のコミュニティが含まれている．そして，グラフ番号 $k-1$ のコミュニティを構成するノード数に

比べて，グラフ番号 k のコミュニティを構成するノード数が増えている．(コミュニティを構成するノードの全てが一致しているとは限らない)

（3） 縮小: コミュニティクラスタにグラフ番号 k のコミュニティとグラフ番号 $k-1$ のコミュニティが含まれている．そして，グラフ番号 $k-1$ のコミュニティを構成するノード数に比べて，グラフ番号 k のコミュニティを構成するノード数が減っている．(コミュニティを構成するノードの全てが一致しているとは限らない)

（4） 維持: コミュニティクラスタにグラフ番号 k のコミュニティとグラフ番号 $k-1$ のコミュニティが含まれている．そして，グラフ番号 $k-1$ のコミュニティを構成するノード数とグラフ番号 k のコミュニティを構成するノード数が等しい．(コミュニティを構成するノードの全てが一致しているとは限らない)

（5） 消滅: コミュニティクラスタにグラフ番号 k のコミュニティが含まれていて，グラフ番号が k より大きいコミュニティが含まれていない．

（6） (維持): コミュニティクラスタにグラフ番号 k のコミュニティが含まれていないが，グラフ番号 k より小さいコミュニティとグラフ番号 k より大きいコミュニティが含まれている．(実際には存在するコミュニティがグラフ番号 k 上で見えなくなっている)

（7） 再現: コミュニティクラスタにグラフ番号 k のコミュニティが含まれていて，グラフ番号 $k-1$ のコミュニティが含まれていないが，グラフ番号が $k-1$ より小さいコミュニティが含まれている．

例えば，あるコミュニティクラスタ $Cl \in \{C_1^5, C_4^6, C_3^9\}$ に対して，上記の「コミュニティの変化」と照合することで抽出できる変化ルールは，

ex) 生成 → 縮小 → (維持) → (維持) → 再現 → 消滅

となる．これは,「グラフ番号 5 で生成されたコミュニティが，次のグラフで縮小した．そして，その後続く 2 つのグラフで見えなくなったが，グラフ番号 9 で再登場し，その次のグラフで消滅した.」という変化ルールを示している．以上のプロセスによって，このような不定期に出現する (系列の途中で見えなくなる) コミュニティの変化ルールが抽出される．

4. 評価実験

本実験では，2 種類の人工データと実際のデータストリームに対して提案手法を適用し，その性能を評価する．1 つ目の人工データセットで，2. 節で述べた関連研究が解析対象とする「時刻 $t-1$ のコミュニティ構造から劇的に逸脱しない」コミュニティの変化を捉えられるかを検証し，2 つ目の人工データセットで本研究が解析対象とする不定期に出現するコミュニティの変化を捉えられるかを検証する．また，実際のデータストリームとして SBM データを解析対象とし，抽出される変化ルールを検証する．

4.1 人工データセット 1

本節では，関連研究 [8], [9] で行われている人工データセットによる性能評価を参考に,「時刻 $t-1$ のコミュニティ構造から劇的に逸脱しない」コミュニティから成るグラフ系列で提案アルゴリズムの性能を評価する．このグラフ系列は，Newman らの研究 [4] で記述された方法で作られたグラフを基に，以下の手順で作成する．

まず，それぞれ 32 個のノードを持つ 4 つのコミュニティで構成されるグラフを作成する．グラフ中のリンクは，平均次数が 16，コミュニティ間に存在するリンクの平均数が 3 となるようにランダムに配置される．ここで，同じコミュニティ中のノードペアでリンクが配置される確率を p_{in}，異なるコミュニティのノードペアでリンクが配置される確率を p_{out} とすると，p_{in} は p_{out} と比べて非常に大きな値となる ($p_{in} \gg p_{out}$)．そして，そのグラフを徐々に変化させた 10 個のグラフから成るグラフ系列を作成する．

これらのパラメータによって生成されるグラフ，及びそれを変化させたグラフの集合であるグラフ系列は，コミュニティ間を繋ぐリンクが少ないことから，CNM 法によってコミュニティ構造が簡単に検出できる人工データである．また，グラフ系列の作成に用いた上記のパラメータは，関連研究 [8], [9] で高い識別性能を示した値を基に設定している．これらのことから本実験は，コミュニティの変化を解析する拡張グラフカーネルと階層的クラスタリングを組み合わせた手法によって，関連研究が解析対象とする「時刻 $t-1$ のコミュニティ構造から劇的に逸脱しない」コミュニティの変化を捉えられるかを検証できると考える．ここで，本実験で用いる各パラメータを表 1 で示す．

表 1　人工データセット 1 のパラメータ設定

	ノードの変化率	λ	CT
評価 1-1	0.1	0.5	1〜10
評価 1-2	0.2	0.5	1〜10
評価 1-3	0.3	0.5	1〜10
評価 1-4	0.4	0.5	1〜10

表 1 中のノードの変化率とは，各コミュニティで変化するノードの割合を意味する．例えば，ノードの変化率が 0.1 のとき，各コミュニティから 3 つのノードがランダムに選択され，それぞれ他の違うコミュニティのランダムに選択されたノードの 1 つとコミュニティ所属を組み換える．また，コミュニティ所属が組み換えられたノードのリンクは，組み換え前のノードが持っていたリンク情報を引き継ぐこととする．そして，抽出されるコミュニティの変化ルールと正解ルールを比較することで，再現率 (式 10 参照) を算出する．

$$\text{再現率} = \frac{\text{抽出された正解ルール}}{\text{正解ルールの総数}} \tag{10}$$

つまり，この実験では，抽出されるコミュニティの変化ルールから算出される再現率が高いほど，提案アルゴリズムが「時刻 $t-1$ のコミュニティ構造から劇的に逸脱しない」コミュニティの変化を捉えられたと解釈できる．ここで，評価結果を図 4 に示す．

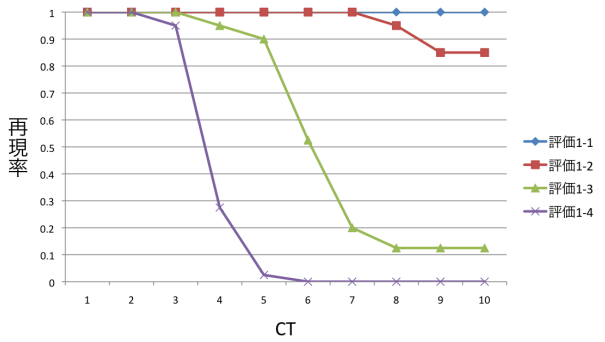

図 4　人工データセット 1 に対する評価結果

図 4 は，各ノードの変化率に対して 10 種類のグラフ系列を作成し，各グラフ系列から抽出された変化ルールを基に算出される再現率の平均値を表している．図 4 に示すように，コミュニティの変化を解析する閾値 CT が 1，2 のとき，4 種類全ての評価結果が変化ルールを正確に抽出できていることを示している．しかし，閾値 CT が 3 以上になると，ノードの変化率が 0.2 以上の評価結果 (1-2〜4) において，徐々に再現率が減少することが示された．特に，ノードの変化率が高いほどより小さな閾値 CT で再現率の減少が顕著に表れている．このことから，「時刻 $t-1$ のコミュニティ構造から劇的に逸脱しない」コミュニティから成るグラフ系列に提案アルゴリズムを適用する場合，閾値 CT が低い設定の場合は関連研究 [8], [9] と同様に高い識別性能を示し，閾値 CT が高い場合には識別性能が低下することが示された．

4.2　人工データセット 2

本節では，不定期に出現する (系列の途中で見えなくなる) ようなコミュニティを含んだ人工データセットで提案アルゴリズムの性能を評価する．ここで用いる人工データセットは，4.1 節で作成されたグラフ系列を基に，それらを大規模グラフ系列に拡張したものである．以下にその作成手順を示す．

まず，それぞれ 32 個のノードを持つ 12 個のコミュニティで構成されるグラフを作成する．グラフ中のリンクは，平均次数が 16，p_{in} と p_{out} の値が 4.1 節で作成されたグラフの値と近くなるように配置される．そして，そのグラフを徐々に変化させた (ノードの変化率を 0.1 と固定)，30 個のグラフから成るグラフ系列を作成する．

これらのパラメータによって生成されるグラフ，及びそれを変化させたグラフの集合であるグラフ系列は，4.1 節のグラフ系列と同様にコミュニティ間を繋ぐリンクが少ないことから，CNM 法によってコミュニティ構造が簡単に検出できる人工データである．そこで，このグラフ系列中の各グラフでランダムにコミュニティを選出し，選出されたコミュニティに該当するノードだけが存在するグラフに書き換える．この人工データに提案アルゴリズムを適用することで，本研究が解析対象とする「不定期に出現する」コミュニティの変化を捉えられるかを検証する．ここで，本実験で用いる各パラメータを表 2 で示す．

表 2 中のコミュニティの出現率とは，各グラフでコミュニティが出現する確率を意味する．例えば，コミュニティの出現率が 0.8 のとき，系列中の各グラフで 12 個のコミュニティそれぞれ

表 2　人工データセット 2 のパラメータ設定
(ノード変化率が 0.1 の場合)

	コミュニティの出現率	λ	CT
評価 2-1	0.8	0.5	1〜30
評価 2-2	0.6	0.5	1〜30
評価 2-3	0.4	0.5	1〜30
評価 2-4	0.2	0.5	1〜30

に乱数 $R(0 \leqq R < 1)$ を割り当て，$R < 0.8$ を満たすコミュニティに該当するノードだけが存在するグラフに書き換える．また，コミュニティの出現率を低く設定すると，グラフ系列中でどのコミュニティも出現しないグラフが作成される可能性があるため，本実験では系列中の各グラフで最低 1 つ以上のコミュニティが存在するように調整した．そして，4.1 節と同様に抽出されるコミュニティの変化ルールと正解ルールを比較することで，再現率 (式 10 参照) を算出する．つまり，この実験では，抽出されるコミュニティの変化ルールから算出される再現率が高いほど，提案アルゴリズムが不定期に出現するコミュニティの変化を捉えられたと解釈できる．ここで，評価結果を図 5 に示す．

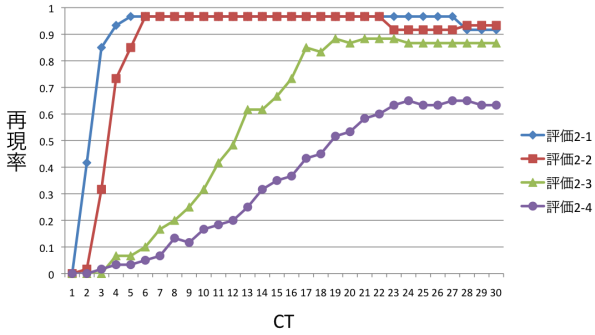

図 5　人工データセット 2 に対する評価結果

図 5 は，それぞれのパラメータ設定に対して 5 種類のグラフ系列を作成し，各グラフ系列から抽出された変化ルールを基に算出される再現率の平均値を表している．図 5 に示すように，連続するグラフ間でしかコミュニティの関係を解析しない $CT = 1$ の場合，どの出現率においても，正解ルールを一つも抽出できなかった．これに対して，閾値 CT を高く設定することで，再現率が高くなることが示された．特に，コミュニティの出現率が高い場合には，より低い閾値 CT で再現率を限りなく 1.0 に近い値まで上昇させることが示された．これらのことから，「不定期に出現する」コミュニティが含まれたグラフ系列に提案アルゴリズムを適用する場合，コミュニティの出現率に応じた閾値 CT を設定することで，不定期に出現するコミュニティの変化を捉えられることが示された．

5.　livedoor clip のデータセット

本節では，実際のデータストリームとして SBM データを用いて提案アルゴリズムを評価する．ここで用いる SBM データは，livedoor 社が提供している livedoor クリップの研究用データセット [15] である．このデータには，ユーザ ID，ブックマークページの URL，ブックマーク作成時刻，登録タグが含まれており，livedoor クリップがサービスを開始した 2006 年 6 月か

ら 2008 年 9 月までの，約 2 万 5 千ユーザの約 150 万件のデータをその解析対象とした．本実験では，実際の SBM データに提案アルゴリズムを適用することによってどのようなコミュニティの変化ルールが抽出されるかを検証する．特に，コミュニティの変化を解析する閾値 CT を 2 種類（$CT = 1$ と $CT = 5$）適用し，抽出される変化ルールにどのように影響を与えるのかを評価する．ここで，$CT = 1$ の評価は，連続する二つのグラフにおけるコミュニティの関係しか解析しないことを意味し，本研究がこれまでに試みた手法 [14] のアプローチと同一のものとなる．

ここで，実際の SBM では正解となる変化ルールが明らかではないため，人工データセットで用いた再現率で評価することができない．そこで，抽出される変化ルールを評価するために，「ステップ数」という評価軸を定義する．この評価軸は，コミュニティの生成を確認されたグラフ番号から消滅を確認されたグラフ番号までの差を表す．例えば，あるコミュニティクラスタ $Cl \in \{C_1^5, C_4^6, C_3^9\}$ から抽出される変化ルールが以下のようなルールであったとき，そのステップ数は 5 となる．

ex) 生成 → 縮小 → (維持) → (維持) → 再現 → 消滅

ここで，上記の変化ルールに含まれる「(維持)」とは，該当するグラフ上ではその存在を確認できなかったコミュニティが，それ以前とそれ以降のグラフで存在を確認された際に，「実際には存在するがそのグラフ上で見えなくなっている」と解釈された結果を意味する．この例の場合，コミュニティクラスタに含まれるコミュニティのグラフ番号は 5，6，9 であり，6 番目と 9 番目のグラフでコミュニティの存在が確認されたことから，コミュニティの存在が確認されなかった 7,8 番目のグラフ上では，そのコミュニティは見えなくなっていたと解釈され「(維持)」という変化の過程が与えられる．

つまり，連続するする二つのグラフにおけるコミュニティの関係しか解析しない $CT = 1$ で抽出されるルールと比べて，$CT = 5$ で抽出されるルールに，より多くのステップ数の多いルールが含まれていることによって，不定期に出現するコミュニティの変化を捉えたと解釈でき，提案手法の有効性を評価できると考える．また，本実験で用いる各パラメータを表 3 で示す．

表 3　SBM データに対するパラメータ設定

	BT	λ	CT
評価 3-1	2500	0.5	1
評価 3-2	2500	0.5	5

本実験における，グラフ系列を作成する際の閾値 BT は，SBM ユーザのブックマーク行動を同じグラフ上で表現するため，試行錯誤的に $BT = 2500$ とした．そして，提案手法によって，121 個のグラフから成るグラフ系列が作成された．また，CNM 法によって抽出されたコミュニティは合計で 3963 個あり，その中でもノード数が 100 以上のコミュニティ，226 個を対象にして，変化ルールを抽出した．解析対象のコミュニティを限定したのは，事前研究 [14] においてノード数が 100 以上の大きなコミュニティの変化を捉えられなかったことを動機として，より大きなコミュニティの変化ルールの抽出を目指しているためである．ここで，評価 3-1，評価 3-2 によって抽出された変化ルールをステップ数ごとにまとめた結果を図 6 に示す．

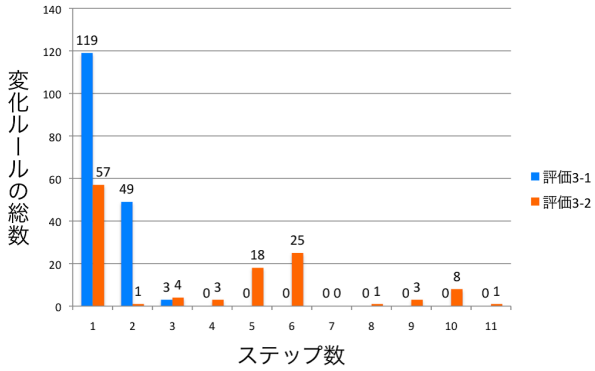

図 6　抽出された変化ルール

図 6 の縦軸は，各ステップ数に該当する変化ルールの種類数ではなく，合計数を表している．例えば，ステップ数が 1 の変化ルールは「生成 → 消滅」の 1 種類しか存在しないが，提案手法を適用した結果，評価 3-1 でそのルールが 119 個抽出されたことを意味する．図 6 に示すように，$CT = 5$ である評価 3-2 は，$CT = 1$ である評価 3-1 に比べて，よりステップ数の多い変化ルールを抽出することが示された．特に評価 3-2 では，ステップ数が 1 の極端に短い変化ルールの数を評価 3-1 の半数以下に抑え，より長いステップの変化ルールを抽出することができた．これらのことから，グラフ系列全体でコミュニティの関係を解析する提案アルゴリズムによって，不定期に出現するようなコミュニティの関係を発見できたと考えられる．そして，事前研究 [14] では発見できなかった，より大きなコミュニティのより長い変化ルールを抽出できることが示された．ここで，解析対象としたコミュニティはノード数が 100 以上のコミュニティである．そのような大規模なコミュニティは簡単に生成されないため，抽出された長い変化ルールにノイズが含まれる可能性は低いと考える。また，抽出された変化ルールの中から特徴的なものとして，ステップ数 5 のルールを表 4 に示す．また，この変化ルールに含まれる期間で，これらのコミュニティを構成するユーザがブックマーク時に付与したタグ情報を頻度順に整理すると，表 5 のようになる．ここで，表 5 中の「アルバイト (859)」とは，その期間でアルバイトというタグが 859 個のトランザクションで登録されたことを意味する．

このタグ情報の結果から，「アルバイト」や「求人」というタグがそれ以外のタグに比べて，圧倒的に多く登録されていることがわかる．特に，「アルバイト」と「求人」のタグを使用しているユーザが，各グラフの中で次数の高いユーザに該当することから，この変化ルールは求職情報を集めているユーザを中心としたコミュニティの変化構造であると解釈できる．また，「アルバイト，求人」をタグとして登録したトランザクション数を日別に集計すると，その上位 5 件は表 6 のようになった．

表 4，表 6 に示すように，解析対象のデータ全体で「アルバイト」「求人」といったタグがブックマーク情報として登録さ

表 4　抽出されたコミュニティの変化ルール例

ステップ数	5
時期	2008-05-30 05:00:00 〜 2008-06-06 05:00:00
変化ルール [グラフ番号]: ノード数	生成 [80]: 280 → (維持)[81]: . . . → 再現 [82]: 142 → 縮小 [83]: 137 → 拡大 [84]: 147 → 消滅 [85]: 0

表 5　抽出されたコミュニティのタグ情報

順位＼グラフ番号	80	82	83	84
1	アルバイト (859)	アルバイト (303)	アルバイト (306)	アルバイト (402)
2	求人 (859)	求人 (303)	求人 (306)	求人 (402)
3	yuiseki(136)	あとで読む (59)	社会 (41)	ネタ (47)
4	社会 (129)	ネタ (42)	mobile(41)	iphone(42)
5	ネタ (113)	社会 (28)	news(33) softbank(33)	web(40)

れた日が集中した数日間を，この変化ルールを基に発見することができたことがわかる．これらの結果から，本研究の提案アルゴリズムによって不定期に出現するコミュニティの変化を捉えただけではなく，大規模なデータに隠れた特徴的なデータ構造を発見できたと考える．

表 6 「アルバイト，求人」タグが登録された日の上位 5 件

順位	日付	件数
1	2008-06-01	774
2	2008-06-05	559
3	2008-06-02	558
4	2008-07-02	483
5	2008-06-03	454

6. ま と め

本研究では，データストリームのトランザクション集合を，ある期間ごとのデータ関係構造とその変化を表すグラフ系列と捉え，そのグラフ系列中のコミュニティの変化を解析するアルゴリズムを提案した．特に，拡張グラフカーネルと階層的クラスタリングを組み合わせた手法によって，コミュニティの変化を解析する関連研究で扱ってこなかった，不定期に出現する (系列の途中で見えなくなる) コミュニティの変化を捉える手法を導入した．そして，人工データセットを用いた実験では,「時刻 $t-1$ のコミュニティ構造から劇的に逸脱しない」コミュニティと「不定期に出現する (系列の途中で見えなくなる)」コミュニティ，それぞれの変化を捉えられることが示された．さらに，実際のソーシャルブックマークデータを用いた評価実験では，不定期に出現するコミュニティの変化を捉えるだけではなく，大規模なデータに隠れた特徴的なデータ構造を発見できることが示された．

今後の課題として，抽出される変化ルールの解釈とその応用が挙げられる．評価実験で用いた SBM データでは，表 5 に示すように求職情報を主たる話題としたコミュニティ構造とその構造の変化を発見できたが，その話題自体の変化は見られなかった．さらに，他に抽出されたルールに関しても，ステップ数の多いルールを中心にその主たる話題 (タグ属性，URL 属性から取得した Web ページのタイトルなど) とその変化を調べたが，解釈の難しいコミュニティばかりであった．そのため，抽出されたコミュニティやそれらの変化ルールに対して効率的にラベリング (意味付け) する手法を確立することで，提案アルゴリズムが様々なデータストリームから特徴的なデータ構造や，それらのコミュニティが持つ話題の変化を発見できると考える．

文　献

[1] A.Clauset, M.E.J. Newman, and C. Moore: Finding community structure in very large networks, Physical Review E, Vol.70, p.066111, 2004.

[2] H. Kashima, K. Tsuda, A. Inokuchi: Marginalized Kernels for Labeled Graphs, *Proceedings of the 20th International Conference on Machine Learning*, pp.321-328, 2003.

[3] J. Shawe-Taylor, N. Cristianini: Kernel Methods for Pattern Analysis, *Cambridge University Press*, 2004.

[4] M. E. J. Newman and M. Girvan: Finding and evaluating community structure in networks, PHYSICAL REVIEW E 69, 026113, 2004.

[5] M. Girvan and M. Newman: Community structure in social and biological networks, *Proceedings of the National Academy of the United States of America*, 99(12), pp.7821-7826, 2002.

[6] M. Gupta, C. Aggarwal, J. Han, Y. Sun: Evolutionary Clustering and Analysis of Heterogeneous Information Networks, IBM Research Report, RC25012(W1006-064) June 17, 2010.

[7] T. Yamaguchi, A. Niimi: Time-Series Analysis Communities using Adaptive Graph Kernels in Data Streams, *Proceedings of Joint 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on advanced Intelligent Systems (SCIS* **&** *ISIS2010)*, 2010.

[8] Y. Chi, X. Song, D. Zhou, K. Hino and B. L. Tseng: On Evolutionary Spectral Clustering, ACM Transactions on Knowledge Discovery from Data, Vol.3, No.4, Article 17, 2009.

[9] Y. Lin, Y. Chi, S. Zhu, H. Sundaram and B. L. Tseng: FacetNet: a framework for analyzing communities and their evolutions in dynamic networks, Proceeding of the 17th international conference on World Wide Web, pp. 685-694, 2008.

[10] 有村博紀: 大規模データストリームのためのマイニング技術の動向, 電子情報通信学会論文誌, D-I J88-D-I(3), pp.563-575, 2005.

[11] 都築学, 小西修: 大規模データストリームのための履歴情報を用いたカーネル法の拡張, 情報処理学会論文誌 データベース, Vol.1, No.3, pp.49-59, 2008.

[12] 豊田正史, 喜連川優: 日本におけるウェブコミュニティの発展過程, 日本データベース学会 letters Vol.2, No.1, pp.35-38, 2003.

[13] 新美礼彦, 小西修: ストリームカーネルマシンによるパラレルブースティング, 情報処理学会論文誌 データベース, Vol.2 No. 4 pp.13–23, 2009.

[14] 山口雄大, 新美礼彦: データストリームに対する相関ルールを用いたコミュニティの時系列解析, 第 24 回人工知能学会全国大会, 2010.

[15] EDGE Datasets, `http://labs.edge.jp/datasets/` (参照 2011-01-23).