対象グラフ集合の特性を反映した構造類似性とクラスタリングへの応用

和田　貴久† 　大野　博之†† 　稲積　宏誠††

† 青山学院大学大学院理工学研究科理工学専攻知能情報コース　〒 229-8558 神奈川県相模原市淵野辺 5-10-1
†† 青山学院大学理工学部情報テクノロジー学科　〒 229-8558 神奈川県相模原市淵野辺 5-10-1
E-mail: †{t-wada,oono,hiro}@ina-lab.it.aoyama.ac.jp

あらまし　蓄積されるデータの多様化に伴い，複雑な構造を持つグラフデータは，ますます増加してきている．そのため，構造データを取り扱うための有用な DB システムやデータマイニング手法の開発は，データの有効活用のためには必須である．そこで，我々は対象とするグラフ集合より特徴的な部分構造を用いることによって定義できる構造類似性を提案し，それに基づくクラスタリングを行ない，特性を評価する．グラフ集合からの特徴的な構造の抽出では，Cl-GBI(Chunkingless Graph Based Induction) を使用する．Cl-GBI は，サポート値や繰り返し数，同時チャンク数などのパラメータによって探索空間を制御できるうえに，探索空間内では部分構造を漏れなく抽出できるため，必要最小限の構造情報を効率的に利用できる．ここで，各グラフの特徴表現としてノード情報を用いる．各ノードは，グラフの構成要素であると同時に部分構造の構成要素でもあるので，その関係を提案する構造分布行列で表現し，グラフ間類似度をその構造分布行列のマッチングと重み付け計算より定義する．以上の手法の特性の検討および応用について考察する．

キーワード　構造類似性, Cl-GBI(Chunkingless Graph Based Induction), グラフマイニング, クラスタリング

New Structure Similarity and Graph Clustering in Accordance with the Feature of Target Graph Sets

Takahisa WADA†, Hiroyuki OONO††, and Hiroshige INAZUMI††

† Graduate school of Science and Engineering, Aoyama Gakuin University　Fuchinobe 5-10-1, Sagamihara-shi, Kanagawa, 229-8558 Japan
†† College of Science and Engineering, Aoyama Gakuin University　Fuchinobe 5-10-1, Sagamihara-shi, Kanagawa, 229-8558 Japan
E-mail: †{t-wada,oono,hiro}@ina-lab.it.aoyama.ac.jp

Abstract　The graph data with the complex structure increases more and more along with the diversification of the accumulated data. Therefore, effectively to leverage data, the development of a useful DB system and the data mining technique to handle structural data is imperative. In this paper, we propose new structure similarity in accordance with the feature of target graph sets, and discusses a graph clustering method based on its criterion. In the extraction of a feature structure from the graph sets, Cl-GBI(Chunkingless Graph Based Induction) is used. Because Cl-GBI can control the search space according to parameters of a support value, a number of repetitions, and simultaneous number etc. of chunks, and a partial structure can be extracted without omission in the search space, structural information on the minimum requirement can be efficiently used. As a feature expression in each graph, each node information is used. The examination and the application of the characteristic of the above-mentioned technique are considered.

Key words　structure similarity, Cl-GBI(Chunkingless Graph Based Induction), graph mining, clustering

1. はじめに

近年，Web 上にはテキストデータや時系列データ，グラフデータなど，さまざまな形式のデータが蓄積され，それらを活用しようと，多くのマイニング手法が研究されている．ただし，従来の解析対象の多くはテキストデータや時系列データな

どのデータそのものであり，構造情報を含むグラフデータに対するマイニング手法の研究は比較的新しい分野といえる．しかしながら，近年このようなグラフデータを扱うマイニング手法については，精力的に研究が進められ，多くの有用なアルゴリズムが提案されている [2]．特に，対象とするグラフ集合に共通する部分グラフ抽出のためのアルゴリズムを用いて，知識発見のための多くの取り組みがなされている．われわれも，特に Graph-Based Induction（GBI）法 [3], [4], [13] に注目して，化学物質の特性を分析するためのツールの開発や，GBI 法により得られた部分グラフ情報を用いた分類問題やクラスタリングへの応用について検討を行ってきた．これは，GBI 法の部分グラフ発見過程の説明容易性やさまざまな制御戦略の適用可能性の高さからであった．

しかしながら，GBI 法ではノードの置き換えと Greedy 探索を基本とするがゆえに，多くの見落としが存在することになる．それを補うために，われわれは，GBI 法を多段的に実行する環境を実現するなどしたが，これはアプリケーション依存のシステムであり，本質的な改良とはいえなかった．しかし，Chunkingless Graph-Based Induction（Cl-GBI）法 [5], [6] が提案されたことによって，従来の GBI 法の欠点は完全に補われることとなり，その応用範囲も広くなったと考えられる．

本稿では，グラフ構造情報のより有効な分析を実現するために，グラフにおける構造上の類似性に注目し，分析対象とするグラフ集合全体の持つ構造上の特徴に基づくグラフ間の構造類似性を検討する．さらに，それを用いた取り組みの典型例として，グラフクラスタリングの問題を取り扱う．すなわち，グラフ構造データから Cl-GBI 法を用いて部分グラフの抽出を行い，それを用いたグラフ間類似度の計算方法を提案し，クラスタリングへと発展させる．グラフ構造は汎用的なデータ構造であり，本稿においても一般的なグラフ構造データに適用できる手法の開発を目的とする．ただし，最も典型的なグラフ構造データとして化学物質を取り上げ，いくつかの実験によって，本稿で提案した構造類似性の特性の検討を行い，その応用について展望する．

2. 構造的類似性の考え方

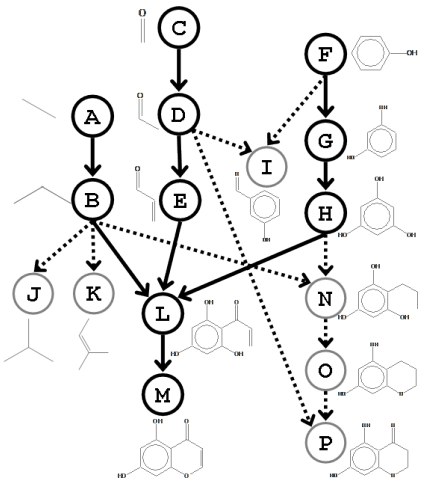
2.1 背　　景

構造的な類似性の活用においては，与えられたグラフ集合のグラフごとに，グラフ中に含まれる連結部分グラフを列挙し，それを数値化したものを利用するという考え方が多く用いられている [8]．その際，部分グラフの特徴をどのように捉えるか，またその特徴に基づいてどのようにクラスタリングを行うかにより，クラスタリング結果が大きく左右されることになる．連結部分グラフの数値化は，ノードあるいはリンクに注目し，それらのもつ情報をどのようの評価するかによって決まる．また，ラベルつきのノード，リンクを扱う際には，さらにその取り扱いが複雑になる．グラフ構造の構造的特徴づけに対しては，化学物質の構造的類似性の観点から TFS （Topological Fragment Spectra）が提案されている [11]．TFS では，ノードの次数の概念を拡張し，連結部分グラフを擬似ノードとみなして，それぞれ次数を定義し，その次数分布を数値化することによって，グラフ構造を多次元空間のベクトルに変換する．これをグラフスペクトルと呼び，多次元ベクトルとみなすことによって，ベクトル間の類似度に基づいたクラスタリングを行うというものである．すなわち，グラフスペクトルそのものが，連結性や密度という観点から，対象とするグラフごとの特性を表現しているとみなすものである．また，グラフスペクトルを基にした分析では，どのようなグラフ集合を対象としているかによらず，個々のグラフの客観的な特徴として定義できるという利点がある．さらに，スペクトル間の類似度の改良や最適なクラスタ数の推定，TFS に化学構造データ固有の情報である骨格プロフィールなどを加味した類似性評価へと拡張するなどの試みもなされている [12]．しかしながら，評価対象とするグラフ集合の中で個々のグラフ間の相対的な類似関係を評価するという観点からは，まず対象とするグラフ集合全体の特徴を自動的に取り込みつつ，その中で類似性評価を実現するような，さらなる検討が望まれるのではないかと考えられる．

2.2 部分構造関係グラフによる構造的類似性

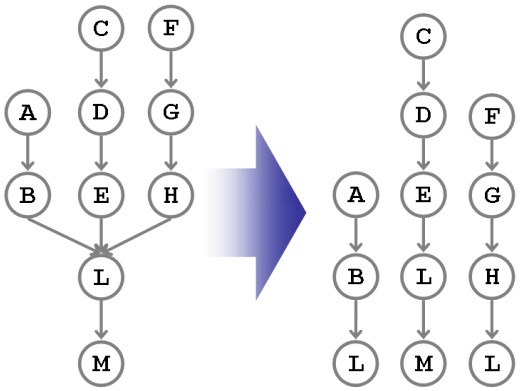
構造的類似性を定義するうえで，われわれは，まず，次のような考え方に注目した．抽出された部分グラフ間に包含関係を導入することによって，定義される部分グラフ間の半順序構造に注目する方法である．この場合には，各グラフが，どのような部分グラフ集合に対応しているかということが，部分グラフの全体集合における半順序関係を保持したまま定義されることになる．従って，類似度評価を行う際には，対象とする 2 つのグラフを表現する部分グラフ集合の共通集合と差集合を，部分グラフの半順序関係を保持したまま定義できるため，それにより特徴の違いを数値化することが可能となる．たとえば，各部分グラフを，それを包含する最大構造を持つ部分グラフ群をもとにしてカテゴライズすることにより，各カテゴリごとに対象とする 2 つのグラフ間の差異を評価することができる．その際，部分グラフの大きさ（たとえばノードの数やリンクの数）をもとにして数値化することなども考えられる．これによって，構成する部分グラフの相対的な関係を類似性評価に反映させることが可能となる．これを部分構造関係グラフによる構造的類異性と呼ぶ．この考え方にもとづく類似度計算法を示す．

まず，グラフ構造データから抽出される全ての部分グラフの包含関係に注目することにより，抽出された部分グラフをノード, 包含関係による半順序関係をリンクとする有向グラフで表現する．ここで，推移的に成り立つ順序関係などの冗長な順序関係は削除し，得られた有向グラフを部分構造関係グラフと呼ぶ．これは，対象とするグラフ集合全体の性質を表わすものであり，任意のグラフは，この部分構造関係グラフの部分グラフに対応することになる．図 1 に部分構造関係グラフの例を示す．これは，部分構造間の半順序関係を表しており，末端ノード，すなわち，その部分構造を包含する他の部分構造がないノードを特定することができる．各末端ノードからたどれるリンクをすべてたどり，部分グラフを抽出する．たとえば，図 2(a) は，図 1 の末端ノード M から求められる部分グラフである．このように，部分グラフ間の包含関係を木構造表現することができ，



部分構造関係グラフ

図 1　部分構造関係グラフの例



(a) 部分構造関係木　　(b) 部分構造関係系列

図 2　部分構造関係群と系列表現の例

部分構造関係グラフは，部分的な重複を許して多くの木構造に分解されることになる．さらに，この木構造を分割することで，図 2(b) のような系列に変換できる．これを部分構造関係系列と呼ぶ．ここで，対象とするグラフ構造データは，各部分構造関係系列を属性とし，そのグラフ構造データに含まれる最大の部分グラフを特定し，それぞれに対して，その部分グラフを属性値とすることで表現できる．その系列に該当する部分グラフを持たない場合には，Φを意味するダミーノードを付加することにより，全ての属性について，属性値が与えられるようにする．各データ間の類似性は，各系列ごとの属性値の共通性から評価を行う．属性値が一致するときは，同一の性質を持つと解釈し，属性値が異なるときは，相違度を計算する．相違度は，属性値のサイズと他の系列で出現している頻度により重み付けされた値の差から求められる．すべての系列の相違度の累計値をデータ間の距離とし，類似度に変換することで類似性の尺度とする．

しかしながら，以上のような部分グラフ集合の半順序構造のみに注目した方法では，重複して保持する部分グラフの個数や部分グラフ間の相対的位置関係を示す情報が欠けているために，本稿の目的とする類似性評価には，不十分であると考えた．

2.3　対象グラフ集合の特性を反映した構造類似性

本稿で提案する手法は，対象とするグラフ集合全体の性質を背景とし，そのなかでの相対的な類似性に基づく分析を実現することを目的とするものである．そのためには，まず対象とするグラフ集合全体を特徴づけることが必要であり，それは，グラフ集合全体のなかにある一定基準以上存在する部分グラフの組合せを評価することにより実現できると考えた．これは，各グラフは，どのような部分グラフをどのように保持しているかにより特徴づけられるという考え方に基づくものである．以下，この前提にたって構造的類似性を評価する．ただし，対象とするデータサイズや個別の性質に大きく左右され過ぎないために，ある程度汎用性の高い部分グラフの組み合わせや，それらの相互関係や相対関係を反映させるものが好ましいと考えた．本稿での考え方は，対象とするグラフ全体をあらわす特徴として，部分グラフ間の包含関係に基づく部分グラフ集合を用いるのではなく，包含関係を有する部分グラフが，対象とするグラフ中にどのように分布しているかを評価する．すなわち，グラフを構成する一つ一つのノードが，どのような部分グラフにどの程度含まれているかを情報として保持し，それを用いてグラフ間の構造的な類似性を評価するという考え方である．そのためには，抽出漏れがなく，効率的に対象とするグラフ集合に共通する部分グラフを抽出する必要がある．また，その際には，同一部分グラフが対象グラフ中に複数存在する場合には，重複を許してすべての部分グラフが抽出される必要がある．そこで，部分グラフ抽出アルゴリズムとしては，Cl-GBI 法を採用することとした．Cl-GBI 法の特徴は，頻度あるいはその他の基準を満足する部分グラフを，重複を許すことで漏れなく抽出することが可能であることに加えて，一定条件を満足するもののなかで，戦略的に優先度をつけて抽出することも可能であるからである [7], [9], [10]．このことは，対象とする集合の特性を反映させることを目的とした本稿の取り組みにおいては，大量に抽出される部分グラフの選別という点で特に有効となる．

3.　部分構造情報に基づく構造類似性

3.1　Cl-GBI 法を用いた部分構造抽出

ノードとリンクで表現されるグラフ構造データは，Cl-GBI 法を適用することによって，高い頻度で出現する特徴的な部分グラフを高速に抽出することができる．

Cl-GBI 法では，まず，グラフ内に存在する 2 つのノードとそれらを結ぶリンクから構成されるノードペアをすべて特定する．その中から出現頻度の高いノードペアをあらかじめ設定するビーム幅分だけ擬似チャンクを行う．擬似チャンクとは，GBI 法で行なわれるチャンク（ノードペアを表現する擬似ノードを生成し，置き換える操作）とは異なり，ノードペアを新規ノードとして完全に置き換えるのではなく，ノードペアのノードとリンクの情報はそのまま保持する形で追加する操作である．この擬似ノードも含めて頻度の高いノードペアの抽出を繰り返し実行する．また，ビーム幅を大きくすることは多様なノードペアや大きなサイズのノードペアを早期に抽出させるなど，効率的なランダムサーチを行うことにも効果的である．この点が，

オリジナルの GBI 法からの最大の改良点であり，これにより，部分的に重なる部分グラフなどのすべての部分グラフを抽出することができることになる．

しかし，本稿で提案する構造類似性を，意味のあるものにするためには，部分グラフの全探索を実現するよりも，初期の段階からサイズの比較的大きいノードペアを抽出することが重要となる．なぜならば，抽出部分グラフの増大は計算量の増大につながるため，なるべく効率的にグラフ間の特徴の違いを表現することのできる部分グラフを抽出しておきたいためである．したがって，ここでは，単純に頻度の大きい順に擬似チャンクを行わずに，一定頻度以上という条件の下にランダム性を持たせることを考える．そこで，擬似チャンクの候補になっているノードペアを，それに含まれている実体ノード数に応じてグループ分けを行い，このグループ数はビーム幅と同一としたうえで，グループごとに最も頻度の高いノードペアを擬似チャンク対象とする．あるいは，頻度順に，一定間隔を置いたうえで，ビーム幅に相当するノードペアを擬似チャンク対象とするなどの方法を採用することとした．これによって，一定頻度を保ったうえで，網羅性を加味しながらノード数の多い部分グラフを比較的早い段階から抽出できることになる．もちろん，この他領域知識に基づく制御戦略を導入することも有効であると考えられる．

次に，抽出された部分グラフがどのグラフに含まれているかについてのチェックを行う．その際，ある部分グラフを含む対象グラフ集合とその部分グラフに包含される部分グラフを含む対象グラフ集合が同一であるならば，後者は，グラフを区別するための構造上の特徴としては冗長なものであり，意味をなさないとみなすことにする．したがって，このような場合には，大きいほうの部分グラフ，すなわち前者のみを保持し，それに含まれる部分グラフである後者は棄却する．本提案手法では，このようにして，なるべく冗長な部分グラフを除去し，かつ多様で網羅性の高い部分グラフを早期に抽出したうえで，構造類似性を評価する．

以上の結果，抽出された部分グラフは，Cl-GBI 法の特性から，チャンク過程を示す履歴情報を用いて，それぞれの包含関係を示す情報が保持されている．また，分析対象であるグラフの各ノードには，擬似ノード生成過程から，抽出されたどの部分グラフの構成要素となっているかという情報が保持されている．これらが，構造類似性を評価するうえでの基本情報となる．

3.2 グラフの構造分布行列表現

Cl-GBI 法により抽出された部分グラフを用いることによって，各グラフは，それを構成するノードと部分グラフとの関係に基づいて表現することができる．ノードラベル集合を L，グラフ G_k のノード n_i^k の総数を I_k とすると，ノード集合 N^k は次のように表すことができる．

$$N^k = \{n_i^k | n_i^k \in L, i = 1, 2, \cdots I_k\} \tag{1}$$

また，対象とするグラフ集合からリンクラベルも考慮して抽出された部分グラフ集合を P，抽出された部分グラフ数を J，ノード n_i^k を含む部分グラフ $p_j \in P, j = 1, 2, \cdots, J$，の数を

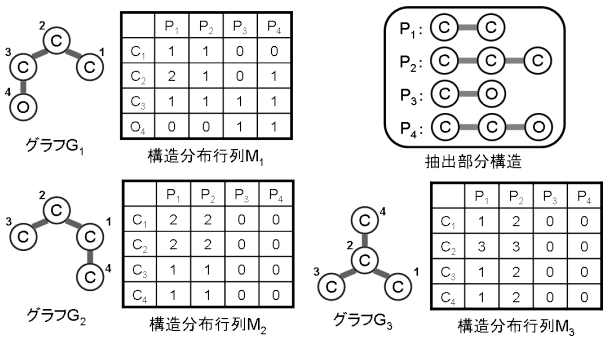


図 3　構造分布行列の例

$m_{ij}^k = m^k(n_i^k, p_j)$ とおくと，任意のグラフ G_k は以下に示す行列 M_k で表現することができる．

$$M_k = \begin{bmatrix} m_{11}^k & m_{12}^k & \cdots & m_{1J}^k \\ m_{21}^k & m_{22}^k & & \\ \vdots & & \ddots & \vdots \\ m_{I_k 1}^k & & \cdots & m_{I_k J}^k \end{bmatrix} \tag{2}$$

この行列 M_k を構造分布行列と呼び，G_k の構造上の特徴が表現されているとみなす．これは，Cl-GBI 法の実行過程の情報から容易に作成することができる．

図 3 に構造分布行列の例を示す．これは，対象とするグラフ集合全体を化学物質としたときに，抽出された部分グラフが p_1：C-C，p_2：C-C-C，p_3：C-O，p_4：C-C-O であったときの，グラフ G_1，G_2，G_3 の特徴表現である．

3.3 グラフ間類似度の算出法

本稿では，まず，任意の 2 つのグラフ中に含まれる共通ラベルを持つノード集合ごとに，各ノード間の類似性を定義し，次に，そのノード間の類似性を用いてグラフ間の構造類似性を定義する．その際，各ノード間の類似性は，関連している部分グラフの共通数の多いノードペアから評価していくこととする．そのため，対象とするグラフ間で同一ラベルを持つノード数が異なる場合には，余分なノードは評価対象としない．たとえば，比較対象となる一方のグラフにのみ，あるノードラベルを持つノードが多数存在しても，その多くのノードは構造類似性の尺度には，直接は反映しない考え方である．しかし，一方にしか含まれないノードは，共通するノードと関連を持つ部分グラフの情報によって間接的に反映される．

まず，任意のグラフ対に対して，構造類似性を表す尺度であるノード間類似度を定義する．いま，比較対象となるグラフを G_1，G_2 とする．簡単のために，ノードラベルを 1 種類，ノード数については，G_1 のほうが G_2 より多いとする．また，対象とするグラフ集合全体から抽出されている部分グラフの数を J，それぞれの構造分布行列を M_1，M_2，さらに，部分グラフ p_i を構成するノード数を $size(p_i)$ とする．このとき，グラフ G_1 のノード x とグラフ G_2 のノード y のノード間類似値 r_{xy}^{12} およびノード間相異値 d_{xy}^{12} を以下のように定義する．

$$r_{xy}^{12} = \sum_{j=1}^{J} \alpha_j \min(m_{xj}^1, m_{yj}^2) \tag{3}$$

$$1 \leqq \alpha_j \leqq size(p_j)$$

$$d_{xy}^{12} = \sum_{j=1}^{J} \beta_j |m_{xj}^1 - m_{yj}^2| \tag{4}$$

$$1 \leqq \beta_j \leqq \alpha_j$$

$$m_{ij}^1 = m^1(n_i^1, p_j) \in M_1$$

$$i = 1, 2, \cdots, I_1$$

$$m_{ij}^2 = m^2(n_i^2, p_j) \in M_2$$

$$i = 1, 2, \cdots, I_2, \quad I_2 \leqq I_1$$

これらを用いて，ノード間類似度 s_{xy}^{12} を以下のように定義する．

$$s_{xy}^{12} = \frac{r_{xy}^{12}}{r_{xy}^{12} + d_{xy}^{12}} \tag{5}$$

ここで，定義されたノード間類似度がより高くなるようなノードペアを選び出す．そのための準備として，2 つのグラフに含まれる部分グラフのうち最も大きな共通部分グラフを用意する．その最大共通部分グラフと関連を持つノードを各グラフより取り出す．そして，それらのノード群の中でノード間類似度を計算し，その値がより高くなるノードペアを見つける．続いて，残りのノード群の中でノード間類似度が高くなるノードペアを見つける．これは，比較するグラフに共通する最も大きな部分グラフに関連するノード同士は，より高い類似度を得る可能性が高いからである．最終的に I_2 個のノードペアを選び，それをグラフ間類似度を評価するための比較ノードペアとして確定する．その結果，グラフ G_1 と G_2 から選択されたノードペア $(x_1, y_1), (x_2, y_2), \cdots, (x_{I_2}, y_{I_2})$ に対して，グラフ G_1 と G_2 のグラフ間類似値 R^{12}, グラフ間相異値 D^{12} およびグラフ間類似度 S^{12} を以下のように定義する．

$$S^{12} = \frac{R^{12}}{R^{12} + D^{12}} \tag{6}$$

$$R^{12} = \sum_{i=1}^{I_2} r_{x_i y_i}^{12} \tag{7}$$

$$D^{12} = \sum_{i=1}^{I_2} d_{x_i y_i}^{12} \tag{8}$$

このようにして定義したグラフ間類似度は，次のような考え方に基づいている．

（1） 各ノードの類似性を評価する際には，同一の部分グラフをもつ個数を加味して評価する．したがって，部分グラフが対象とするグラフ中にどのように分布しているかという点についても，類似性評価に加味されることになる．

（2） 部分グラフのサイズを類似性評価に加味することができることとしている．これは，大きな部分グラフを共有するかしないかという点が，類似性の評価に大きく影響するであろうとの考え方による．ただし，類似値については，$1 \leqq \alpha_j \leqq size(p_j)$，相異値については，$1 \leqq \beta_j \leqq \alpha_j$ とする．これは，部分グラフのサイズの反映については，類似値が相異値を下回らないという条件を置いたことによる．

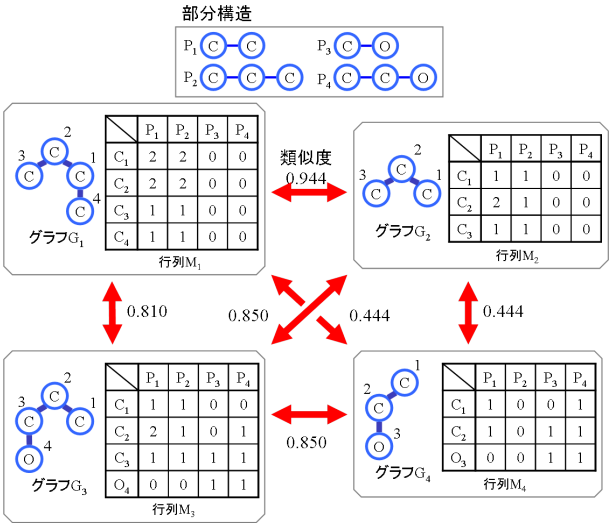


図 4　構造分布行列とグラフ間類似度

3.4　グラフ間類似度の特徴

本稿で提案するグラフ間類似度では，特に，ノード数が大きく異なる場合や，一方のグラフのみにノードラベルが存在し，他方にはそのラベルが存在しない場合に，適切な類似度評価が行えているかどうかについての検討が必要となる．

そこで，ここでも化学物質を例として，図 2 に示す 4 つのグラフを用いて考えてみることとする．ここでは，ノード間類似値および相異値の重みは $\alpha_j = size(p_j)$,$\beta_j = 1$ とし，類似部分を高く評価することとする．リンクラベルは 1 種類，ノードラベルは，C と O の 2 種類とし，グラフ G_1 とグラフ G_3 がノード数が 4 で一方に O が含まれ，グラフ G_2 とグラフ G_4 がノード数 3 で一方に O が含まれている．また，抽出されている部分構造も図中にあるとおり，少数ラベルである O についても，部分グラフとしての頻度は基準以上であり，評価対象となることを意味している．4 つのグラフ間のグラフ間類似度では，同様の条件ながら，$S^{13} \geqq S^{24}$ となっている．まず，グラフを構成するノード数が少ないほど，類似度を評価する際のノード数の差やラベルの違いによる影響が反映されることが推測される．

図 5 と図 6 には，ノード間類似度とグラフ間類似度をまとめた．5 を見ると，G_2 は，G_1，G_3 に同様に包含されているにもかかわらず，$S^{12} \geqq S^{23}$ となっていることがわかる．このことは，ノードラベルの違いが，抽出された部分グラフに反映される場合には，ノードラベルの違いにより類似度を下げるように機能することがわかる．しかし，部分グラフである P_3 や P_4 が一定頻度に満たず，採択されない場合には，それらが，相異値として反映されなくなるため，$S'^{23} = 1$，$S'^{12} = 0.944$ となり，$S'^{12} \leqq S'^{23}$ と逆転することになる．このように，ある共通部分を共有したうえで，異なるラベルが付加されているような場合には，その付加ラベルを含む部分グラフがどのような頻度であるかによって，グラフ間類似度が制御されることになる．次に 6 を見ると，G_2 と G_4 が G_3 に包含されており，$S^{23} = S^{24}$ となっている．しかし，この場合でも，もし部分グラフである p_3 や p_4 が一定頻度に満たず，採択されない場合には，それ

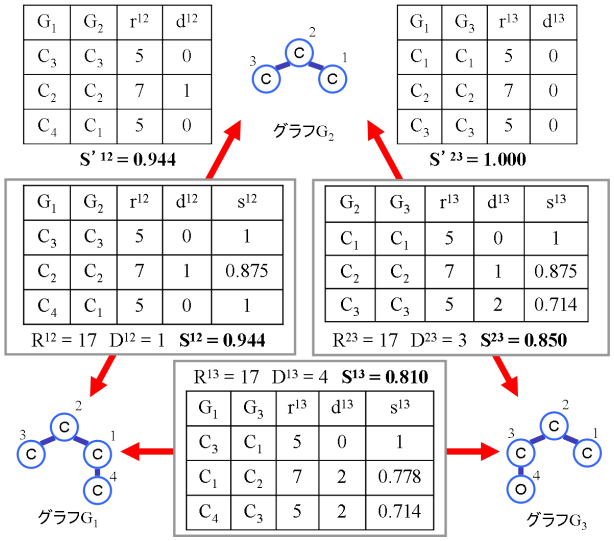


図 5　グラフ間類似度の抽出部分グラフによる影響 1

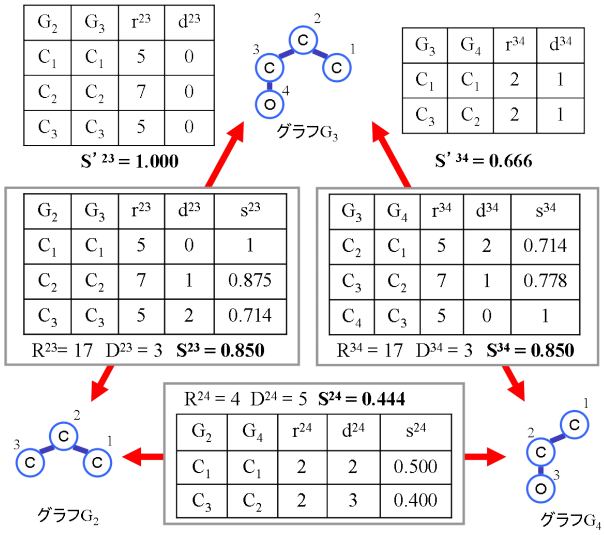


図 6　グラフ間類似度の抽出部分グラフによる影響 2

らが，相異値として反映されなくなるため，$S'^{23} = 1.000$ となる一方で，$S'^{34} = 0.666$ となり，$S'^{34} \leqq S'^{23}$ となる．また，$S'^{24} = 0.666$ となり，ここでは，グラフ間類似度が増加することがわかる．ここでも，評価対象となるグラフ集合のもつ全体としての性質が，クラスタリングの基準となるグラフ間類似度に反映されていることがわかる．

したがって，本稿で提案したグラフ間類似度は，対象とする 2 つのグラフ間でのノード数の違いやラベルの偏りについて，次のような特性を示すものと考えられる．

（1） ノード数が大きく異なる場合：少ないノード数を持つグラフを基準にノード間類似度が評価されることになる．しかし，多くのノードをもつグラフにのみ存在する部分グラフが，対象とするグラフ中に一定頻度以上存在し，対象とするグラフ集合のなかで採択されている場合には，一部のノードではその影響を受けて相異値の増大を生むことなり，相対的に類似度を低下させることとなる．しかし，多くのノードをもつグラフにのみ存在する部分グラフが，対象とするグラフ集合にほとんど

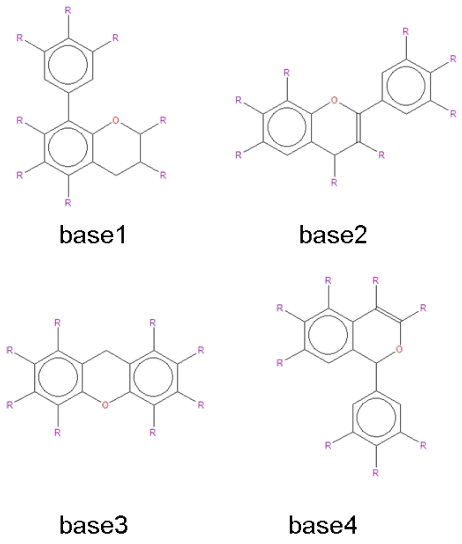


図 7　基盤グラフ

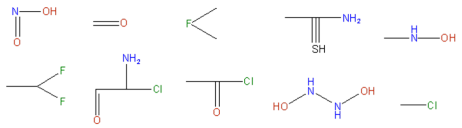


図 8　置換基の例

存在しない場合には，その部分グラフの影響が相異値として反映されないため，類似度を低下させる要因とはならない．

（2） 一方のグラフのみにノードラベルが存在し，他方にはそのラベルが存在しない場合：該当するノードラベルについてのノード間類似度はまったく評価されない．しかし，該当するノードラベルをもつグラフにのみ存在する部分グラフが，対象とするグラフ集合中に一定頻度以上存在しており，その部分グラフが，対象とするグラフに共通するノードラベルとの組み合わせで構成されている場合には，一部のノードで相異値の増大を生むことなり，相対的に類似度を低下させることとなる．しかし，多くのノードをもつグラフにのみ存在する部分グラフが，対象とするグラフ集合にほとんど存在しない場合には，その部分グラフの影響が相異値として反映されないため，類似度を低下させる要因とはならない．

3.5　階層的クラスタリング

クラスタリングアルゴリズムは，目的に応じて用いることを想定している．また，クラスタサイズの決定についても検討を行う必要はあるが，本稿では，提案した類似度に基づく評価の妥当性を検証する目的で，求められたすべてのグラフ間類似度を用いて，最短距離法，最長距離法，群平均法による階層的クラスタリングを行うこととした．作成されたデンドログラムに対して閾値を設定し，複数のクラスタに分割する．これによって，閾値の変化に対して，どのようなクラスタが生成されるかに注目することとした．

4.　実　　験

本稿で提案している構造的類似性の有効性を示すため，部分構造関係グラフによる構造的類似性との間でクラスタリング性

表 1　Cl-GBI 法の条件

	条件 A	条件 B
Level	10	10
Beam 幅	5	5
含有数の閾値	40 個 (20 %)	80 個 (40 %)

表 2　Cl-GBI 法による抽出結果

	条件 A	条件 B
抽出グラフの種類	137 種類	35 種類
抽出グラフの総数	25033 個	11516 個
抽出時間	4 分 43 秒 906	3 分 9 秒 296

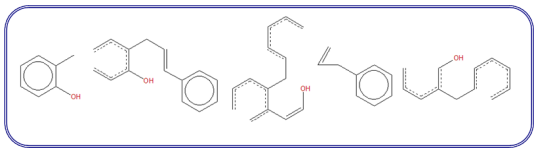


図 9　条件 A の抽出部分グラフの例

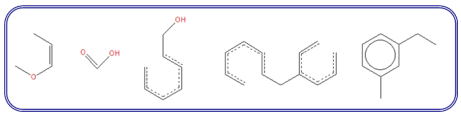


図 10　条件 B の抽出部分グラフの例

能の比較を行う．

4.1　実験データ

実験データとして，図 7 に示す 4 つの仮想化学構造式を基盤グラフとして，各 R に置換基をランダムに配置したものを 50 個ずつ，合計 200 個のグラフを生成した．これらの基盤グラフとして挙げた 4 つの仮想化学構造は，すべてベンゼン環を 2 つずつ含んでおり，ベンゼン環でない環を 1 つずつ含んでいる．しかし，環の配置がすべて異なる．すなわち，構成要素は非常に類似した構造を含んでいるが，結合の仕方によって各々の構造に違いが出ている基盤グラフであるといえる．置換基は，全部で 16 種類あり，ノード数が 1 から 8 までの様々なサイズのものを用意した．一例を図 8 に示す．

4.2　検 証 結 果

ここでは，仮想化学構造と置換基から生成された 200 個のグラフ構造データが，それぞれどの基盤グラフを含んでいるかわからないものと想定してクラスタリングを行う．そして，生成されたクラスタ内でどの基盤グラフを含んでいるデータがいくつあるかを比較することで行う．

まず，仮想化学構造と置換基から生成された 200 個のグラフ構造データに対して，Cl-GBI 法を適用する．表 1，2 に，Cl-GBI 法のパラメータの設定と抽出結果を示す．条件 A では，全体の 40 個以上のグラフに含まれる 137 種類の部分グラフを抽出し，図 9 は，抽出部分グラフの一例を示している．また，条件 B では，全体の 80 個以上のグラフに含まれる 35 種類の部分グラフを抽出し，図 10 に抽出部分グラフの一例を示している．2 つの条件での抽出グラフを比較すると，条件 A では，特定の基盤グラフにのみ含まれる部分グラフが多く抽出されているのに対し，条件 B では，どの基盤グラフにも含まれうるグラフが抽出されている．これらの抽出された部分グラフを用いて，2 手法による類似度計算とクラスタリングを行う．クラスタリングは，階層的クラスタリングである群平均法によりデンドログラムを作成することで行う．

表 3　条件 A で抽出した部分グラフを用いたクラスタリング結果

(a) 提案している類似性

cluster	base1	base2	base3	base4
1	29	0	0	0
2	15	0	0	0
3	3	0	0	0
4	3	0	0	0
5	0	50	0	0
6	0	0	36	0
7	0	0	14	0
8	0	0	0	50

(b) 部分構造関係グラフによる類似性

cluster	base1	base2	base3	base4
1	36	0	0	0
2	14	0	0	0
3	0	50	0	0
4	0	0	28	0
5	0	0	22	0
6	0	0	0	26
7	0	0	0	21
8	0	0	0	3

表 4　条件 B で抽出した部分グラフを用いたクラスタリング結果

(a) 提案している類似性

cluster	base1	base2	base3	base4
1	16	0	0	0
2	34	0	0	0
3	0	20	0	0
4	0	8	0	0
5	0	6	0	0
6	0	5	0	0
7	0	4	0	0
8	0	4	0	0
9	0	3	0	0
10	0	0	50	0
11	0	0	0	50

(b) 部分構造関係グラフによる類似性

cluster	base1	base2	base3	base4
1	32	0	0	0
2	15	0	0	0
3	1	7	0	0
4	2	30	0	0
5	0	13	0	0
6	0	0	21	0
7	0	0	8	0
8	0	0	16	3
9	0	0	5	2
10	0	0	0	35
11	0	0	0	10

条件 A で抽出した部分グラフを用いて行われたクラスタリング結果を表 3 に示す．ここでは，提案している類似性を用いた結果と部分構造関係グラフを用いた結果の両方で，共通する 1 つの基盤グラフを含むデータで構成されるクラスタに分割するクラスタリングを行うことができた．これは，抽出された部分グラフの中に基盤グラフの特徴を表現している部分グラフが多く存在しているので，その部分グラフの有無の情報が大きく反映されているためであると考えられる．一方，条件 B で抽出した部分グラフを用いて行われたクラスタリング結果を表 4 に示す．(a) の提案している類似性を用いたクラスタリング結果は，共通する 1 つの基盤グラフを含むデータで構成されるクラスタに分割しているが，(b) の部分構造関係グラフによる類似性を用いたクラスタリング結果は cluster3，4，8，9 のように，含む基盤グラフが異なるデータが混在するクラスタが生成されている．以上のことから，提案している類似性は基盤グラフそのものを特徴付ける部分グラフが抽出されていなくとも，小さな部分グラフの組み合わせで基盤グラフの特徴を表現できているといえる．

図 11 は，表 4(b) の cluster3 内のグラフ base1_002，base2_100 と cluster5 内のグラフ base2_097 を取り上げ，その類似度の比較を表している．部分構造関係グラフによる類似性は，base1_002 と base2_100 のように含んでいる基盤グラフが異なり，大きく構造上の特徴が違うグラフの類似度が 0.979 と高い値となっており，含んでいる基盤グラフが同じものはそれより低い 0.697 という値になっている．これは，含んでいる部分グラフの共通性にのみ注目しているため，グラフの一部分の配置が異なると

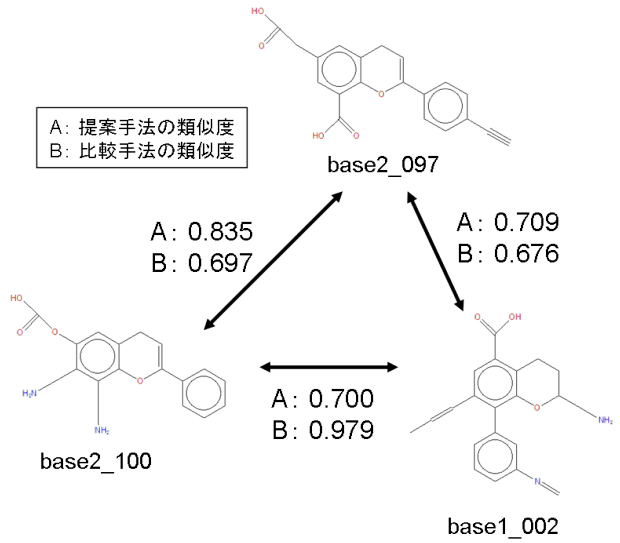


図 11　類似度の比較

いったような構造上の違いをうまく評価できていないといえる．しかし，提案している類似性は，含んでいる部分グラフとノードの関係を考慮しているため，部分グラフのグラフ内の配置を評価でき，グラフそのものの特徴を表現するのに十分大きな部分グラフが抽出されていなくとも，比較的小さな部分グラフの組み合わせから構造上の違いをうまく評価できているといえる．

5. ま と め

本稿で提案した構造類似性を反映したグラフ間類似度の定義は，対象とするグラフ集合の特徴を反映するものとして定義された．これは，グラフを構成する各ノードの特徴を，対象とする集合に存在する部分グラフとどのような関わりを持つかにより定義し，各グラフはそのノードの特徴の集合体とみなすことにより実現される．ただし，これを実現するための部分グラフ抽出については，Cl-GBI 法を用いるのが適当であると考えた．なぜならば，部分グラフを漏れなく抽出することによって，詳細な類似度の定義が実現されるとともに，必要に応じて，粒度の細かなクラスタリングを行うことが可能となることが予想されるからである．これについては，いまだ詳細な議論を行っていないが，現段階では，Cl-GBI 法における頻度の設定によって，共通部分グラフの条件を与え，そのうえで網羅性を維持する部分グラフ抽出を行うことの利点を確認している．この点については，実験により，一定レベルの妥当な結果が得られたが，より詳細な検討を行うことによって，対象とするグラフ集合の性質と分析目的に応じて，必要とする部分グラフのもつ要件を明確にする必要がある．もちろん，これには，領域知識に基づくアプローチも含まれる．

今後は，以上の点について明確な取り組みを行い，より厳密なクラスタリングアルゴリズムとしての確立を図りたい．また，より汎用的な手法としてさまざまな分野への適用も検討していきたい．

文　　献

[1] 速水亜希子, 稲積宏誠：部分構造の包含関係を指標とするグラフクラスタリングの提案－化学物質を対象として－, 人工知能学会知識ベースシステム研究会, SIG-KBS-A405, pp.1-6 (2005)

[2] Kuramochi,M. and Karypis.,G.: An E.cient Algorithm for Discovering Frequent Subgraphs, *IEEE Trans. Knowledge and Data Engineering*, Vol.16, No.9, pp.1038-1051 (2004)

[3] 松田喬, 元田浩, 鷲尾隆：一般グラフ構造データに対する Graph-Based Induction とその応用, 人工知能学会誌, Vol.16, No.4, pp.363-374 (2001).

[4] Matsuda,T., Motoda,H., Yoshida,T. and Washio,T.：Mining Patterns from Structured Data by Beam-wise Graph-Based Induction, *Proc. of DS2002*, pp.422-429 (2002)

[5] Nguyen,P.,Ohara,K.,Motoda,H. and Washio,T. : Cl-GBI : A novel strategy to extract typical patterns from graph data, *SIG-KBS-A403*, pp.105-110 (2004)

[6] Nguyen,P.,Ohara,K.,Motoda,H. and Washio,T. : Cl-GBI : A Novel Approach for Extracting Typical Patterns from Graph-Structured Data., *Proc. of PAKDD2005*, pp.639.649 (2005)

[7] Nguyen,P., Ohara,K., Mogi,A., Motoda,H. and Washio,T. : Constructing Decision Trees for Graph-Structured Data by Chunkingless Graph-Based Induction., *Proc. of PAKDD2006*, pp.390-399 (2006)

[8] Palmer,C., Gibbons,P. and Faloutsos,C. : ANF: A Fast and Scalable Tool for Data Mining in Massive Graphs, *Proc. of the KDD-2002* (2002)

[9] 高林健登, Phu Chien Nguyen, 大原剛三, 元田浩, 鷲尾隆：グラフ構造データからの特徴的なパターン抽出における探索の効率化, 第 19 回人工知能学会全国大会, 2F3-1 (2005)

[10] 高林健登, Phu Chien Nguyen, 大原剛三, 元田浩, 鷲尾隆：グラフ構造データからの特徴的なパターン抽出における制約に基づく探索制御, 第 20 回人工知能学会全国大会, 1A2-4(2006)

[11] Takahashi,Y., Ohoka,H. and Ishiyama,Y.: Structural Similarity Analysis based on Topological Fragement Spectra, *Adavances in Molecular Similarity*, Vol.2, pp.93-104 (1998)

[12] 高橋由雅, 藤島悟志, 加藤博明： 化学物質の構造類似性にもとづくデータマイニング, *J. Comput. Chem. Jpn.*, Vol.2, No.4,pp.119-126 (2003).

[13] 吉田健一, 元田浩：逐次ペア拡張に基づく帰納推論, 人工知能学会誌, Vol.12, No.1, pp.58-97 (1997)