クラスタリング　ニ　モトヅク　ジョウホウ　ノ　ケンサク　ト　シカクカ

堀田, 政ニ

https://doi.org/10.15017/1398256

出版情報：Kyushu Institute of Design, 2001, 博士（工学）, 課程博士
バージョン：
権利関係：

第1章

序論

1.1 研究の目的と背景

データ検索，情報フィルタリングやデータマイニングなど大量の構造的データの探索においてクラスタリングは重要なデータ構造要約法の一つであり [1]，これまでに多くの手法が提案されている [2]．データ構造は，ハイパーテキストなど陽に有向グラフとして表されるものから，文献とキーワードのデータ行列などグラフとは明示的な関係がない多変量データに至るまで，一般に無向あるいは有向グラフとして表すと扱い易くなる．そこでグラフ構造データのクラスタリング法が上記のデータ探索において有用になると思われる．しかし，クラスタリングにグラフ理論的手法を用いるというアプローチは研究されている [3] が，グラフそのものをクラスタリングの対象とした研究は少ないようである [4]．ましてファジークラスタリングに関してはファジー c 平均法 [5] などのようにデータの特徴量に基づく手法がほとんどであり，グラフを対象とした研究はほとんどされていない．

クラスタリングは一般に NP 困難であり，大規模な問題では近似解法が必要とされるが，そのような近似解法の一つとして整数制約を実数に緩和して 2 次計画法，線形計画法あるいは固有値問題に帰着させるアプローチがある．グラフに関する組合せ問題についてそのような緩和により固有値問題に帰着させる一連の近似解法はグラフスペクトル法と総称されている [6]．しかし，グラフスペクトル法でも専らハードクラスタリングが扱われており [7]，ファジークラスタリングに関する研究は少ない．ハードクラスタリングはデータの要約効率は高いがデータ間の細かい位相が失われてしまうので柔軟性に欠ける．特に人間が関わるような曖昧度の高いデータに対してはファジークラスタリングが重要になると思われる．

そこで本論文では，グラフの節点集合をファジークラスタ (ファジー部分グラフ)

に分割するクラスタリング法を提案する．まずはじめに，もっとも単純なグラフ構造である無向グラフからファジークラスタを抽出する方法を示し，これを有向グラフや2部無向グラフに拡張する．これらのクラスタリングは固有値問題に帰着されるので一種のグラフスペクトル法といえる．次に，複数のグラフが混成された複雑なグラフからファジークラスタを抽出する方法を提案する．例えば，ウェブページに関する情報として，ページ間のリンク関係と，各ページのキーワードの出現頻度が与えられたとする．この場合，リンクが有向辺で表され，ページとキーワードとの関係が無向辺で表されるので，全体は有向グラフと2部無向グラフとを合成したものになる．このような複雑なグラフ構造データからファジークラスタを抽出する問題は固有値問題として表すことができない．そこで，複雑なグラフ構造データからのファジークラスタ抽出法として，べき乗法を一般化した反復法による抽出法を導くことにする．

本論文では，上記のファジークラスタリングをデータ検索に応用することを考える．データ検索は，ユーザが入力した問合せ (query) に該当するデータを出力するハンティング検索 (hunting retrieval) と，データを2次元や3次元の低次元空間に表示し，全体を眺めながらユーザ自身が探索を行うブラウジング検索 (browsing retrieval) に大別される．

ハンティング検索では，例えばキーワードに基づくデータ検索の場合，各データに付けられたキーワードをデータ個別に見るのでなく，データ全体に渡る大局的なキーワード付与分布に基づいて検索することが重要である [8]．これは各データのキーワード付与分布の変動が大きく，信頼性が低いためである．大域的な情報抽出法としては，クラスタリング [8] や潜在意味解析法 (Latent Semantic Indexing, LSI)[9](付録 A 参照) などの手法がある．例えば表 1.1 のようなデータ行列が得られたとする [9]．ここで縦の1から9は文献の番号，横の1から12はキーワードの番号であり，要素の0や1や2はその文献 (の説明文) にそのキーワードが登場する回数である．この共起関係行列をそのまま使って，例えばキーワード 10 をクエリとして文献を検索すると文献 6 と 7 と 8 が検索される．しかし，キーワード間の共起関係を見るとわかるように，この場合，文献 9 も検索されるのが妥当である．なぜなら文献 9 はキーワード 10 を直接含んではいないが，キーワード 9, 11, 12 を含んでおり，特にキーワード 11, 12 を含んでいる文献 7, 8 と内容的には類似していると推測できるからである．LSI 法によるハンティング検索では，文献 6, 7, 8 に加えて文献 9 も検索結果に含まれるようになる [9]．一方，クラスタリングに基づくハンティング検索では専らハードクラスタリングが使われており，クラスタ内の個体差の情報は失われるので，クラスタレベルでの粗い検索しか行えなえず，LSI 法のような柔軟な検索ができなかった．また，

表 1.1: データ行列の例

documents	keywords											
	1	2	3	4	5	6	7	8	9	10	11	12
1	1	1	1	0	0	0	0	0	0	0	0	0
2	0	0	1	1	1	1	1	0	1	0	0	0
3	0	1	0	1	1	0	0	1	0	0	0	0
4	1	0	0	0	2	0	0	1	0	0	0	0
5	0	0	0	1	0	1	1	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	1	0	0
7	0	0	0	0	0	0	0	0	0	1	1	0
8	0	0	0	0	0	0	0	0	0	1	1	1
9	0	0	0	0	0	0	0	0	1	0	1	1

データ構造が表 1.1 のような単純なものではなく，ざまざまなグラフが組合さった複雑なデータから大域的な情報を抽出する方法はこれまで提案されておらず，したがって，複雑なグラフ構造データに対する有効なハンティング検索法もなかった．

ブラウジング検索ではデータを 2 次元や 3 次元のユークリッド空間に配置し，データ構造を視覚化することが主な目的となる．ここで視覚化とは，データの構造を視覚的に把握するのを支援するために，数値データを図的に表示することを指す．各データが特徴ベクトルで与えられている場合には，各データの高次元での距離関係をできるだけ保つようにデータを 2 次元や 3 次元空間に写像することが望ましい．このような配置法として多次元尺度構成法 (Multi-Dimensional Scaling, MDS)[10](付録 C 参照) や自己組織化写像 (Self-Organizing Map, SOM)[11](付録 D 参照) が挙げられるが，これらの方法は非線形な反復解法であり，1) 初期値によって結果の配置が変わる，2) 通常ゆっくり収束させる必要があるので長い計算時間を要する，といった問題点がある．いろいろなデータについて配置を比較するときには，このような多数の解を持つ方法では，結果の違いがデータの違いによるのか収束解の違いによるのか判断できないため好ましくない．一方，データが特徴ベクトルとして与えられず，データ間の距離だけしか与えられない場合も多い．この場合，各データをグラフの点，各データの大きさや重要度などを点の重み，データ間の類似度や引用度などを辺の重みとすれば，データは重み付きグラフとして表すことができる．このようなグラ

フの視覚化は与えられたデータ構造の把握に有用であるため，これまでに研究が盛んに行われてきた[12]．しかし，スプリングモデル[13](付録 E 参照) 等で代表されるグラフ自動描画の主目的は視認性や美観であり，データ構造の把握という観点はあまり考慮されていない．

そこで本論文では，グラフ構造データからファジークラスタを抽出し，各データのクラスタへのメンバシップ値を利用してハンティング検索とブラウジング検索を行う方法を提案する．提案手法のハンティング検索では，LSI 法と同様にデータ間の潜在的な関連性に基づいた検索が行える．一方，ブラウジング検索では，各データのクラスタへのメンバシップ値に基づいて，すなわち，クラスタの構造に従ってデータを数量化 3 類で低次元空間に配置することにより，データ構造を反映した配置を高速に行うことが可能である．

更に本論文では，クラスタリングを利用した画像検索の高速化についても提案を行う．画像検索では，大規模な高次元データベースを検索する必要があり，高速化が望まれる．これまでにもクエリに近い k 個の画像を選び出す kNN 検索を高速化する方法として，フィルタリングを利用する方法[14, 15, 16, 17] や，近似的に検索を行うことによって高速化する方法[18] が提案されている．フィルタリングの基本は距離の不等式に基づいて探索範囲を限定することである．距離の不等式としては，三角不等式や次元削減による不等式などがある．次元削減による不等式を使うものとしては，Hafner ら[14] が提案した，カラーヒストグラム間の非類似度を 2 次形式距離 (付録 B 参照) で測り，特徴ベクトルの次元削減に基づくフィルタリングによって領域検索を高速化する方法がある．また，三角不等式を利用する手法としては，Fukunaga ら[19] が提案した，データベースのクラスタリングに基づく分枝限定法によって kNN 検索を高速化する手法がある．本論文では，画像検索の高速化法として，特徴ベクトルの次元削減に基づく距離の不等式とクラスタリングに基づく三角不等式とを組合せた手法と，特徴ベクトルの次元圧縮とクラスタリングを組合せて画像の近似 kNN 検索を行うことにより検索を高速化する方法を提案する．これらの提案手法により従来の方法よりも高速に検索を行うことが可能である．

1.2 論文の概要と構成

第 1 章では，本研究の目的と背景を示し，あわせて論文の概要について述べる．

第 2 章では，類似度行列からファジークラスタを逐次に抽出する方法を提案する．まずはじめに，類似度行列からファジークラスタを抽出する問題が固有値問題に帰着

されることを示す．次に，各データの残存率を類似度行列の要素に掛けて抽出済みのクラスタを取り除くことにより，逐次にクラスタを抽出していく方法を示す．抽出処理はクラスタの凝集度の変化に基づき，ある程度の大きさのクラスタが抽出されたら終了する．提案手法の有効性を，簡単な2次元データに対する実験とグレイスケール画像のセグメンテーション，およびビデオのセグメンテーションの実験例で示す．

第3章では，グラフの隣接行列に基づいて点集合をファジークラスタ(ファジー部分グラフ)に分割する方法を提案する．まずはじめに，点と辺に重みを持つ無向グラフに第2章で提案したファジークラスタ逐次抽出法を適用し，続いてそれを有向グラフや2部無向グラフに拡張してグラフ構造データからファジークラスタを逐次に抽出する方法を提案する．簡単な比較実験により本手法の有効性を示す．

第4章では，第3章のファジークラスタ逐次抽出法を，複数のグラフが混成された複雑なグラフに適用できるように拡張する．例えば，ウェブページに関する情報として，ページ間のリンク関係と，各ページのキーワードの出現頻度が与えられたとする．この場合，リンクが有向辺で表され，ページとキーワードとの関係が無向辺で表されるので，全体は有向グラフと2部無向グラフとを合成したものになる．このような複雑なグラフ構造データには第3章のクラスタリング法は適用できない．そこで，このようなグラフ構造データからファジークラスタを抽出する方法として，べき乗法を一般化した反復法による抽出法を導く．はじめに，無向グラフと2部無向グラフ，および有向グラフから反復法に基づいてファジークラスタを逐次に抽出する方法を示す．次に，有向グラフと2部無向グラフが組合わさったグラフからファジークラスタを抽出する方法を示す．また，これらの反復法の収束性を証明する．最後に，簡単な実験例を示し提案手法の有効性を検証する．

第5章では，クラスタリングに基づいて数量化理論によってデータ構造を視覚化する方法を提案する．数量化理論として数量化3類と数量化4類をとりあげ，数量化4類を重み付き数量化4類に拡張し，固有値問題に帰着できることを示す．続いて数量化3類と数量化4類についてデータ行列に基づく視覚化に関する比較実験を行い，両者の長所と短所を明らかにする．さらに，クラスタリングの結果を重み付き数量化4類により逐次展開して表示する方法を提案し実験例を示す．

第6章では，ファジークラスタリングに基づくハンティング検索法とブラウジング検索法を提案する．まずはじめに，ハンティング検索について簡単な例題を使ってLSI法と比較を行いながら説明する．次に，グラフ構造データの表示法として，第5章の数量化3類による視覚化法をグラフデータに応用し，メンバシップ値を第3軸とする3次元表示によって各データ点の主要度も把握できるように拡張した方法を提案

する．いくつかの具体的な例題により提案手法の有効性を示す．

第7章では，クエリに近いk個の画像を選び出すkNN検索を高速化する手法として，次元削減とクラスタリングに基づくフィルタリングによって高速化する方法を提案する．ここで用いるクラスタリング法はk平均法である．まずはじめに，Hafnerらによって提案された2次形式距離と距離の不等式に基づくフィルタリングによってkNN検索の計算時間を短縮する方法を説明する．このフィルタリングをk平均法に適用することにより，クラスタリングの計算時間が短縮されることを実験で示す．続いてクラスタリングと三角不等式を利用するFukunagaらのフィルタリング法を概説し，特徴ベクトルの次元削減とクラスタリングとを組合せたフィルタリングを用いてkNN検索の計算量を削減する方法を提案する．実験によりkNN検索がフィルタリングによって高速化されることを示す．

第8章では，特徴ベクトルの次元圧縮とクラスタリングを組合せて画像の近似kNN検索を行うことにより検索を高速化する方法を提案する．提案手法では，はじめに画像間の非類似度をカラーヒストグラム間の2次形式距離で測り，主成分分析によってヒストグラムを次元圧縮する．次に，データベース画像を低次元で予めクラスタリングしておき，クエリが入力されたらクエリに近いクラスタを何個か選び，選ばれたクラスタに含まれる画像全体のなかからクエリに近いものをk個選んで出力する．この方法により，フィルタリングを用いる第7章の方法よりも更に高速に検索が行えることを実験で示す．

最後に第9章では，本研究のまとめと今後の課題を示す．

なお，本論文中の実験では，計算機としてDELL Dimension 4100（OS Windows 2000, CPU Pentium-III 1GHz, Memory 383MB）を使用し，計算時間はすべて実行時間で計測している．また，付録として本研究と関連が深い研究内容を巻末に掲載しておく．