ファジークラスタリングによる視覚化と検索のグラフ構造データへの応用

堀田政二[†] 井上光平[†] 浦浜喜一[†]

筆者らは以前，類似度データや共起関係データをファジークラスタリングする手法を提案した．またそのファジークラスタリングで得られるメンバシップに基づいて数量化3類でデータを視覚化する方法も提案した．本論文ではこのデータ視覚化法をグラフ構造データに応用してグラフ構造を視覚化する手法を示す．クラスタリング法としては，無向グラフや2部無向グラフには類似度データや共起関係データでの方法が使えることを示し，有向グラフは2部無向グラフに帰着できることを示す．このクラスタリングの視覚化以外の応用例としてハンティング検索やウェブリンクの推薦などにも利用できることを示す．さらにグラフ構造が複雑化した場合として上記の基本グラフが結合した例にもファジークラスタリング法を拡張し，キーワードによるウェブページのハンティング検索やブラウジング検索へ応用する．

Visualization and Retrieval of Graph-structured Data by Fuzzy Clustering

Seiji Hotta,[†] Kohei Inoue[†] and Kiichi Urahama[†]

We have presented a method for fuzzy clustering of similarity data and that of co-occurrence relational data. We have also presented a data visualization method based on the memberships obtained by the fuzzy clustering. In the present paper, we apply this data visualization method to graph-structured data for visualization of graphical structure of those data. It is shown at first that the clustering method for similarity data and that for relational data can be used for undirected graphs and bipartite ones, and directed graphs can be reduced to bipartite graphs. Besides the application to data visualization, the clustering is available for hunting retrieval of data and recommendation of web links. The clustering method is further extended to more complex graphs composed of some elementary ones, and the extended method is applied to hunting retrieval of web pages by keywords and their browsing.

1. まえがき

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1つである[1]．各データを点とし，データ間の関係を辺で表せばデータ集合はグラフで表現され，データの構造が理解しやすくなる．そこでグラフに対する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り，ウェブ[2]やプログラム[3]のクラスタリングなどが研究されている．このような通常行われるクラスタリングは組合せ的であり，一般にNP困難であるが，曖昧性の高いデータではそのようなハードクラスタリングでなく，整数制約条件を実数に緩和して得られるファジークラスタリングが有用となる．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[4]．筆者らは以前に，ファジークラスタを逐次抽出するグラフスペクトル法の1種を提案し[5]，ウェブの解析[6]やデータ検索[7]に応用した．また，得られたメンバシップに基づいて数量化3類でデータを視覚化する手法も提案した[8]．本論文ではこの視覚化法をグラフ構造データに拡張し，応用例を示す．またさらに進んでファジークラスタリング法を複数のグラフが混成された複雑なグラフに拡張し，その応用例をあげる．

まず最初に，単純なグラフ構造データの例として無向グラフで表される場合について考え，類似度データとの等価性から，前に提案したファジークラスタリング法[5]が適用できることを述べ，後の展開の基礎となる式を示し，同様に前に提案した類似度データの数量化3類による視覚化法[8]が適用できることを述べ，その手順を示す．次に2部無向グラフについても共起関

† 九州芸術工科大学画像設計学科
Faculty of Visual Communication Design, Kyushu Institute of Design

係データとの等価性から，前に提案した共起関係データのファジークラスタリング法[7]が適用できることを述べ，その手順を示す．次に有向グラフについて，2部無向グラフに帰着できることを述べ，数量化3類による視覚化を応用し，ウェブページの例で実験を行い，視覚化以外の応用例としてリンクの推薦を提案する．ここで推薦とは，リンクを張った方がよいと思われるいくつかのページを各ページに対して推薦することをいう．以上のグラフ構造データのファジークラスタリングには前に提案した類似度データや共起関係データの手法が適用できたが，これらの手法が適用できない場合として，グラフ構造が複雑化した例をあげ，そのファジークラスタリング法を導き，リンクとキーワードとによるウェブのクラスタリングに基づく検索と視覚化に応用する．

2. 無向グラフのファジークラスタリングと視覚化

まず最初に最も単純な場合として，データが無向グラフとして表される場合を考える．この場合には辺の重みをデータ間の類似度と見なせば，以前に提案した類似度データのファジークラスタリング法[5]と視覚化法[8]が適用できる．

2.1 ファジークラスタの逐次抽出

m 個の点からなる無向グラフを考え，第 i 点の重みを v_i，第 i 点と第 j 点の間の辺の重みを w_{ij} とする（無向であるから $w_{ij}=w_{ji}$ である）．この点集合からクラスタを順番に取り出す．ここで取り出すという意味は，各点の重みが削減されていくことを表し，点の個数はずっと m 個のままである．第 i 点が第1クラスタに所属する割合を x_{1i} とすると第1クラスタの凝集度は $\sum_{i=1}^{m}\sum_{j=1}^{m} v_i x_{1i} w_{ij} v_j x_{1j}$ で評価され，これが最大となる第1クラスタを

$$\begin{aligned} \max_{x_1} \quad & \sum_{i=1}^{m}\sum_{j=1}^{m} v_i x_{1i} w_{ij} v_j x_{1j} \\ \text{subj.to} \quad & \sum_{i=1}^{m} v_i x_{1i}^2 = 1 \end{aligned} \tag{1}$$

によって求める[5]．式 (1) は $\tilde{x}_{1i}=\sqrt{v_i}x_{1i}$ という変数変換によって固有値問題に帰着し，その第1固有ベクトルを求めることにより解くことができる[5]が，ここでは式 (1) を直接，反復法で解くことにする．これは後の5章での式が固有値問題には帰着されず反復法でしか解けないので，解法を全体で統一するためである．Lagrange 乗数法により式 (1) の解は

$$x_{1i} = \frac{\sum_{j=1}^{m} w_{ij} v_j x_{1j}}{\sqrt{\sum_{i=1}^{m} v_i \left(\sum_{j=1}^{m} w_{ij} v_j x_{1j}\right)^2}} \tag{2}$$

を満たす（付録参照）．そこで $x_1=[x_{11},\ldots,x_{1m}]^T$ を任意に初期設定し，それを式 (2) の右辺に代入して x_1 を更新していく．x_{1i} が最大の i を i_1 とすると $p_{1i}=x_{1i}/x_{1i_1}$ がデータ i の第1クラスタへのメンバシップ値となる[5]．

次に第2クラスタ抽出では第1クラスタを取り除いて同じことを行う．すなわち各点の重みを $(1-p_{1i})v_i$ とする．したがって，各点が第2クラスタに所属する割合 x_{2i} は式 (1) の v_i を $(1-p_{1i})v_i$ に変えて同じ反復法を行えば求まる．第3クラスタ以降も同様であり，一般に第 k クラスタでは式 (1) の v_i を $\prod_{l=1}^{k-1}(1-p_{li})v_i$ として同じことを行えばよい．抽出されるクラスタの凝集度は順番が進むに従って単調に減少する．この凝集度の変化に基づいてある程度まとまったクラスタを抽出し終えたところで抽出を止める．

2.2 数量化3類によるデータの視覚化

以上で得られるクラスタはファジーである．ハードクラスタリングではメンバシップ値が1か0であるから，クラスタ内部でのデータの位置関係の情報は失われてしまうが，ファジークラスタリングではデータの近接関係がメンバシップ値として保持されており，詳細なデータ配置に利用することができる．筆者らは前に，類似度データについて数量化3類によってデータを配置する手法を提案した[8]．上記の無向グラフデータの表示にはこの手法が適用できる．その手順を以下に述べる．

数量化3類は対応分析とも呼ばれ，個体と項目の該当関係がデータ行列として与えられたときに，類似した個体や項目が互いに近くなるように空間に配置する多変量解析法である[9]．個体が m 個，項目が n 個あり，第 j 項目が第 i 個体に該当する頻度が d_{ij} であるとする．このデータ行列 $D=[d_{ij}]$ に基づいて数量化3類で個体を2次元空間に配置するには，$m\le n$ ならば F と G を対角行列：$F=\mathrm{diag}(f_i)$; $f_i=\sum_{j=1}^{n} d_{ij}$，$G=\mathrm{diag}(g_j)$; $g_j=\sum_{i=1}^{m} d_{ij}$ として行列 $F^{-\frac{1}{2}}DG^{-1}D^TF^{-\frac{1}{2}}$ の第2と第3固有ベクトル u_2，u_3 を求めて $x=F^{-\frac{1}{2}}u_2$，$y=F^{-\frac{1}{2}}u_3$ とし，$m>n$ ならば行列 $G^{-\frac{1}{2}}D^TF^{-1}DG^{-\frac{1}{2}}$ の第2と第3固有ベクトル v_2，v_3 と固有値 λ_2，λ_3 を求めて $x=F^{-1}DG^{-\frac{1}{2}}v_2/\sqrt{\lambda_2}$，$y=F^{-1}DG^{-\frac{1}{2}}v_3/\sqrt{\lambda_3}$ と

すれば (x_i, y_i) が第 i 個体の 2 次元座標となる．

この数量化 3 類をファジークラスタリング結果の表示に応用する．個体をグラフの点，項目をクラスタとすると，d_{ij} は各点が各クラスタに該当する頻度であり，メンバシップとなる．すなわち第 i 点の第 j クラスタへのメンバシップ p_{ji} をデータ行列の要素 d_{ij} として数量化 3 類で各点を配置すればメンバシップ値が似ている点は互いに近くに配置され，クラスタ構造が視覚的にとらえやすくなる．上記のように数量化 3 類では個体と項目のうち数が少ない方のサイズの行列の固有値問題を解くだけでよく，今の場合クラスタの数はデータの数よりも少ないので，データの数がいくら多くてもクラスタ数のサイズの固有値問題を解くだけでよく，計算量が少ない．また固有値問題は一意に解が求まるので多次元尺度法[9]，自己組織化ネット[10]，スプリングモデル[11]などのような収束性の問題もない．なお，LSI（Latent Semantic Indexing）法[12]も本章の無向グラフや次の 2 部無向グラフの配置表示に適用できるが，4 章以降のような有向グラフを含む場合には適用できない．

3. 2 部無向グラフのファジークラスタリングと視覚化

次に 2 部無向グラフすなわち点が 2 つの部分集合 $i \in S_1;\ i = 1, \ldots, m$ と $j \in S_2;\ j = 1, \ldots, n$ とに分かれており，部分集合間には重み w_{ij} の無向辺（すなわち $w_{ij} = w_{ji}$）があり，部分集合内には辺がない場合を考える．たとえば文書とキーワードをグラフの点として文書の集合を一方の部分集合 S_1，キーワードの集合をもう一方の部分集合 S_2 として，各キーワードが各文書に登場する回数を辺の重みとすれば 2 部グラフが得られる．たとえば**表 1** のようなデータ行列は**図 1** の無向グラフで表される．逆に 2 部無向グラフはデータ行列で表されるので共起関係データのファジークラスタリング法[7]が適用できる．

3.1 手　　法

部分集合 S_1 をクラスタリングする場合についてクラスタ抽出手順を説明する．部分集合 S_1 の点 i が第 1 クラスタに所属する割合を x_{1i}，S_2 の点 j の所属度を y_{1j} とすると第 1 クラスタは

$$\begin{aligned} \max_{x_1, y_1} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} v_i x_{1i} w_{ij} v_j y_{1j} \\ \text{subj.to} \quad & \sum_{i=1}^{m} v_i x_{1i}^2 = \sum_{j=1}^{n} v_j y_{1j}^2 = 1 \end{aligned} \tag{3}$$

で求められる[7]．これも 2.1 節と同様に $\tilde{x}_{1i} = \sqrt{v_i} x_{1i}$，

表 1　文書とキーワードのデータ行列の例
Table 1　A data matrix for texts and keywords.

	kw1	kw2	kw3	kw4
text1	1	0	1	0
text2	1	1	0	1
text3	0	1	0	1

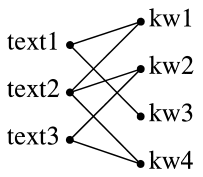


図 1　表 1 のグラフ表現
Fig. 1　Graph representation for Table 1.

$\tilde{y}_{1j} = \sqrt{v_j} y_{1j}$ と変数変換すれば固有値問題に帰着できるが，ここでは式 (3) を直接，反復法で解く．この理由も 2 章と同じく解法を全体で統一するためである．Lagrange 乗数法により式 (3) から

$$x_{1i} = \frac{\sum_{j=1}^{n} w_{ij} v_j y_{1j}}{\sqrt{\sum_{i=1}^{m} v_i \left(\sum_{j=1}^{n} w_{ij} v_j y_{1j} \right)^2}} \tag{4}$$

$$y_{1j} = \frac{\sum_{i=1}^{m} v_i w_{ij} x_{1i}}{\sqrt{\sum_{j=1}^{n} v_j \left(\sum_{i=1}^{m} v_i w_{ij} x_{1i} \right)^2}} \tag{5}$$

が導かれる（付録参照）．そこで $y_1 = [y_{11}, \ldots, y_{1n}]^T$ を任意に初期設定し，それを式 (4) の右辺に代入して x_1 を求め，それを式 (5) の右辺に代入して y_1 を更新していく．S_1 内で x_{1i} が最大の i を i_1 とすると $p_{1i} = x_{1i}/x_{1i_1}$ がデータ $i \in S_1$ の第 1 クラスタへのメンバシップ値となり，S_2 内で y_{1j} が最大の j を j_1 とすると $q_{1j} = y_{1j}/y_{1j_1}$ がデータ $j \in S_2$ の第 1 クラスタへのメンバシップ値となる[7]．

次に第 2 クラスタ抽出では，今は部分集合 S_1 をクラスタリングしているのであるから，点 $i \in S_1$ の重みが $(1 - p_{1i})v_i$ と削減され，点 $j \in S_2$ の重みは v_j のままとする．したがって，各点が第 2 クラスタに所属する割合 x_{2i}，y_{2j} は，式 (3) の v_i を $(1 - p_{1i})v_i$ に変えて同じことをすれば求まる．第 3 クラスタ以降も同様であり，一般に第 k クラスタでは式 (3) の v_i を $\prod_{l=1}^{k-1}(1 - p_{li})v_i$ として同じことを行えばよい．こ

の場合も抽出されるクラスタの凝集度は順番が進むに従って単調に減少するので，この凝集度の変化に基づいてクラスタ数を決める．

なお，この場合クラスタリングされているのは部分集合 S_1 だけであるが，クラスタへのメンバシップは S_1 の点 $i = 1, \ldots, m$ だけでなく，S_2 の点 $j = 1, \ldots, n$ に対しても得られる．たとえば文書をクラスタリングした場合，キーワードの各クラスタへのメンバシップも得られ，各文書に関連するキーワードを知ることができる．また以上では片方の部分集合だけをクラスタリングするとしたが，両部分集合を同時にクラスタリングすることもでき，その場合には両方の部分集合の点の重みを削減していけばよい．

筆者らは前に，このようなファジークラスタリングのメンバシップをハンティング検索に利用する手法を提案した[7]．それはクエリとして入力される文書やキーワードの各クラスタへのメンバシップ q_k; $k = 1, \ldots, N$（N はクラスタ数）を求め，データベース文書 i のメンバシップ $p_{1i}, \ldots, p_{Ni}$ とのコサイン $c_i = \sum_{k=1}^{N} q_k p_{ki} / \sqrt{\sum_{k=1}^{N} q_k^2} \sqrt{\sum_{k=1}^{N} p_{ki}^2}$ を計算して，c_i が大きい文書から選ぶ方法である．複数のキーワードがクエリとして入力される場合には，クエリキーワード j と文書 i とのコサイン c_{ij} を求め，AND なら $s_i = \min_j c_{ij}$ とし，OR なら $s_i = \max_j c_{ij}$ として s_i が大きい文書から順に選ぶ．なお統合法としてこのほかにも AND では $s_i = \prod_j c_{ij}$ とし，OR なら $s_i = \sum_j c_{ij}$ とする方法もあるが，後で示す実験では min と max の方が良好な結果が得られた．

3.2 実 験 例

ここでは前に報告した画像検索への応用の実験結果[7]を再記しておく．そこでは 46 個のキーワードが割り付けられた 160 枚の画像を用いた．画像の物理特徴は用いず，キーワードとの対応関係に基づいて画像をクラスタリングした．本検索法の計算時間は 0.22 秒であり，同じデータに LSI 法を適用した結果 0.33 秒かかった．本方法の方が解析する行列が小さいぶん LSI 法よりも速くなった．一方検索性能は LSI 法とほぼ同等であった．すなわち本方法も LSI 法と同様にクエリキーワードを直接には含まない画像でも意味内容的にクエリに関係があれば検索することができた[7]．

4. 有向グラフのファジークラスタリングと視覚化

以上のように無向グラフや 2 部無向グラフには類似度データ[5]や共起関係データ[7]のクラスタリング法が適用できた．次に有向グラフについて考える．有向グラフでは $w_{ij} \neq w_{ji}$ である．これはこのままでは類似度データや共起関係データには帰着できないが，各点 i を 2 つの点 $\tilde{i}$ と $\hat{i}$ に分けて点の数を 2 倍にし，$\tilde{i}$ から $\hat{j}$ に無向辺を引き，その重みを w_{ij} とすれば 2 部無向グラフに帰着することができる．たとえば**図 2** (a) の有向グラフは図 2 (b) の 2 部無向グラフとなる．

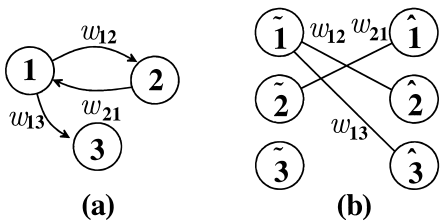


図 2 有向グラフ (a) と等価な 2 部無向グラフ (b)
Fig. 2 Undirected bipartite graph (b) equivalent to digraph (a).

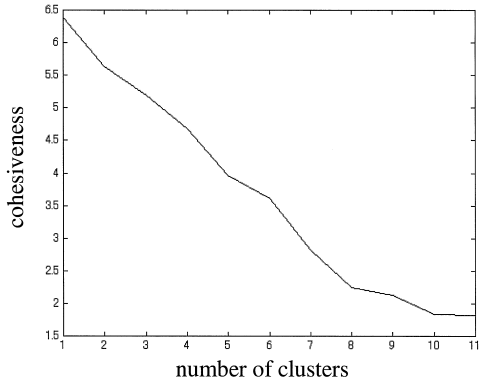


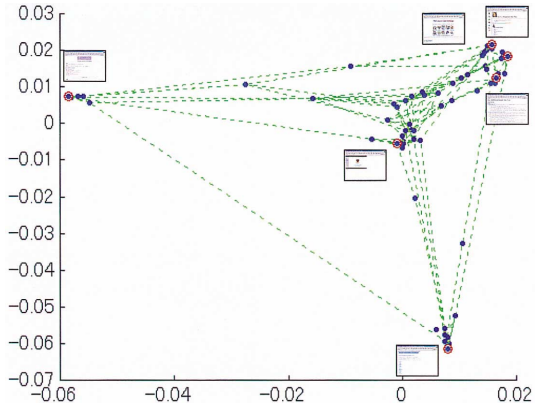
図 3 凝集度の変化
Fig. 3 Variation in cohesiveness of clusters.

4.1 手 法

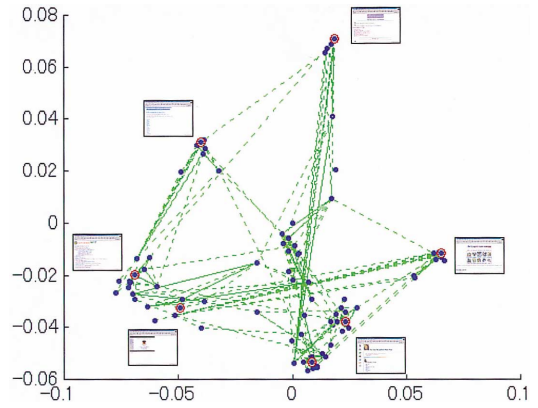
このことから有向グラフは前章の 2 部無向グラフと同じ方法でクラスタリングでき，各点について辺の始点としてのメンバシップ p_i と終点としてのメンバシップ q_j が得られることになる．ただしこの 2 部無向グラフ表現はあくまで形式的であり，$\tilde{i}$ と $\hat{i}$ は元は同じ点 i であるので，これらのメンバシップ p_i と q_j は同時に削減されていく．

4.2 実 験 例

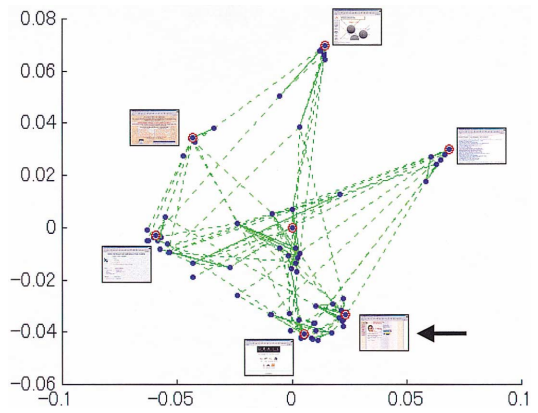
ここでは例としてウェブページのクラスタリングをしてみた．“Pattern Recognition” というキーワードで検索して得られた 118 個のページについて，互いのリンクの関係を調べて有向グラフを構成した．点の重みもリンクの重みもすべて 1 とした．まず最初にリンクの向きは無視して無向グラフとして 2.1 節の方法でクラスタを抽出した．**図 3** に抽出されたクラスタの凝



(a) ignoring link direction

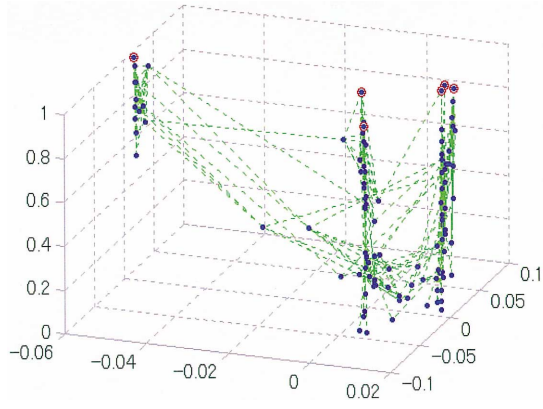


(b) by initial memberships

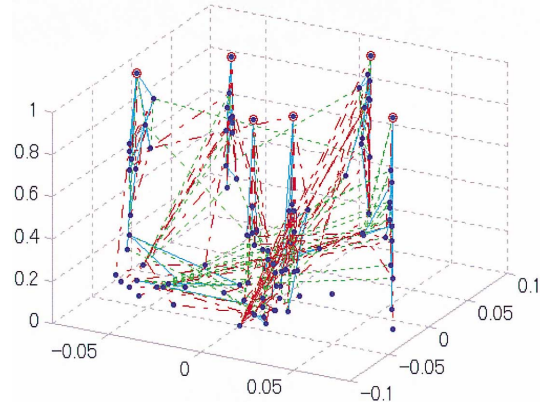


(c) by terminal memberships

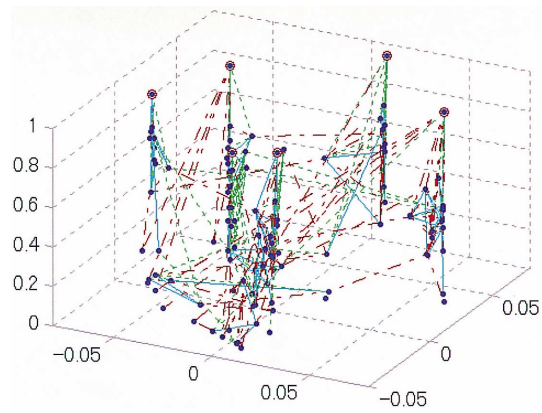
図 4 数量化 3 類によるページの配置
Fig. 4 Arrangement of web pages by the quantification theory 3.



(a) ignoring link direction



(b) by initial memberships



(c) by terminal memberships

図 5 メンバシップを z 座標とした表示
Fig. 5 Display with memberships as z-coordinate.

集度の変化を示す．8 番めのクラスタからは凝集度の減少が比較的緩やかになっており，それ以後はあまりまとまったクラスタは抽出されていない．したがって，ある程度まとまったクラスタは 7 個であることが分かる．得られたメンバシップに基づいて数量化 3 類で各ページを配置した結果を**図 4** (a) に示す．点がページであり，線がリンクを表す．右中央付近のものを除いた 6 個のクラスタの代表ページを大きく表示している．

次に有向グラフとしてクラスタリングを行い，リンクの始点としての各ページのメンバシップに基づいて配置した結果を図 4 (b) に，また終点としてのメンバシップによって配置した結果を図 4 (c) に示す．図 4 (a), (b), (c) とも，だいたい 7 個のクラスタからなることが分かる．図 4 (a) ではリンクの向きによらず多数のリンクに接続するページのメンバシップ値が大きい．また図 4 (b) ではクラスタ内のページの多くにリンクを張っているページがクラスタの代表となる．このようなページを Kleinberg ら[13]はハブと呼んでいる．一方，図 4 (c) では多くのページからリンクを張られているページがクラスタの代表となる．そのようなページはオーソリティと呼ばれる[13]．これらハブやオーソリティは前述の検索と同様に各ページのリンク数をカウントするだけでは検出できない．他のページを介した間接的なリンクも考慮しないといけないからである．メンバシップにはそのような高次の情報が統合されている．これは，ファジークラスタリングを行うときにページ間のリンクを通してメンバシップが計算されているからである．**図 5** は xy 座標を図 4 の配置とし，各ページの最大のメンバシップ値を z 座標としたものである．図 5 (b), (c) の緑の点線は z 座標が小さなページから大きなページへ向かうリンクであり，赤の鎖線は逆の向きのリンクである．空色の実線は相互リンクを表す．部分的に線が重なっているので見にくいが，図 5 (b) では赤の鎖線が多く，図 5 (c) では緑の点線の方が多い．これらの表示からリンクの推薦情報をある程度得ることができる．たとえば**図 6** は図 4 (c) の右下付近のクラスタ（代表ページに矢印を付けた）を拡大したものである．右端の黒四角が代表ページ（オーソリティ）であり，点線はそのページへのリンクを表す．いくつかのページはこのクラスタに所属しているにもかかわらずオーソリティへのリンクを張っていない．そのようなページにとって，オーソリティへのリンクは推薦候補となる．このような視覚的な推薦でなくても 3 章の終わりで述べたハンティング検索と同様な方法で数値的にリンクの推薦を行うこともできる．それには各ページ i の各クラスタ k への始点としてのメンバシップ p_{ki} と終点としてのメンバシップ q_{ki} とから，$c_{ij} = \sum_{k=1}^{N} p_{ki}q_{kj} / \sqrt{\sum_{k=1}^{N} p_{ki}^2} \sqrt{\sum_{k=1}^{N} q_{kj}^2}$ を計算する．これは始点としてのページ i と終点としてのページ j の関連度を表しており，よってページ i からページ j へのリンクの期待度ともいえる．そこで各ページ i について他のページ j への c_{ij} を求め，これが大きい値であるにもかかわらずまだリンクがないようなページに対してはリンクを張ることが推奨される．

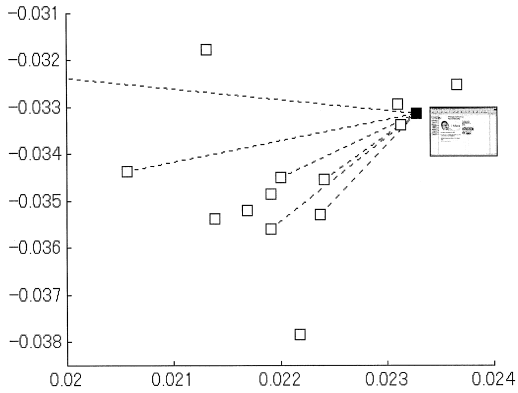


図 6 図 4 (c) の右下付近のクラスタの拡大図
Fig. 6 Magnification of lower right portion in Fig. 4 (c).

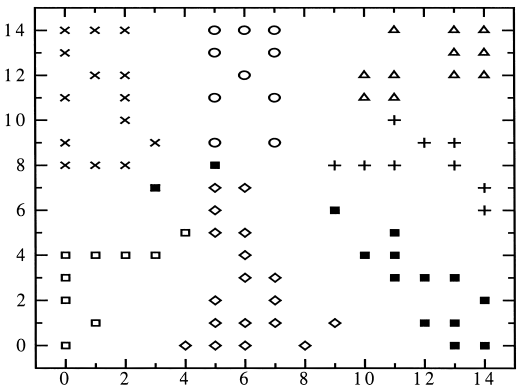


図 7 SOM によるウェブページの配置
Fig. 7 Arrangement of web pages by SOM.

なお，比較のために数量化 4 類，自己組織化ネット（SOM），スプリングモデルでも実験してみた．SOM ではデータが特徴ベクトルとして与えられる必要があるのでメンバシップを特徴ベクトルとして SOM を適用した．計算時間は本方法（数量化 3 類）が 0.01 秒，数量化 4 類も 0.01 秒，SOM が 2.58 秒，スプリングモデルが 18.1 秒であった．本方法の配置結果は図 4 (a) であるが，数量化 4 類による配置ではクラスタの相関関係に歪が生じることが知られている[8]．**図 7** に SOM による配置結果を示す．見やすいようにリンクは省略してある．○ や △ などはクラスタの違い（クラスタは 7 個）を表す．図 7 の配置は図 4 (a) に比べ，データが均等に分布しておりデータは見やすいがクラスタ間の距離関係は歪んでいる．スプリングモデルではクラスタ構造がほとんどつかめないような配置が得られた．

5. 混成グラフのファジークラスタリングと視覚化

以上，無向グラフおよび 2 部無向グラフや有向グラフのクラスタリングとそれによるデータ視覚化を行ってきた．これらの場合にはすでに報告したクラスタリング法[5),7)]を適用することができた．しかし，たとえばウェブのリンクだけでなく各ページのキーワードも与えられる場合には，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，全体は有向グラフと 2 部無向グラフとを合体させたものになる．ここで有向グラフの部分を等価な 2 部無向グラフに変換すると全体は 3 部無向グラフとなる．このような複雑なグラフ構造データには以前のクラスタリング法は適用できない．そこでここでは新たにそのような場合についてのファジークラスタリング法を導く．

5.1 手　　法

例として，向きを無視したリンクとキーワードとに基づいてウェブページをクラスタリングする場合について考えてみる．ページを $i = 1, \ldots, m$ とし，キーワードを $j = 1, \ldots, n$ とする．ページ i からページ i' へのリンクの重み（0 か 1）を $w_{ii'}$（向きは無視するので $w_{ii'} = w_{i'i}$）とし，キーワード j のページ i での頻度を w_{ij} とする．ページ i が第 1 クラスタに所属する割合を x_{1i}，キーワード j の所属度を y_{1j} とすると，式 (1) と (3) を混合した

$$\max_{x_1, y_1} \quad \alpha \sum_{i=1}^{m} \sum_{i'=1}^{m} v_i x_{1i} w_{ii'} v_{i'} x_{1i'} + (1-\alpha) \sum_{i=1}^{m} \sum_{j=1}^{n} v_i x_{1i} w_{ij} v_j y_{1j} \quad (6)$$

$$\text{subj.to} \quad \sum_{i=1}^{m} v_i x_{1i}^2 = 1, \quad \sum_{j=1}^{n} v_j y_{1j}^2 = 1$$

で第 1 クラスタが求められる．$0 \leq \alpha \leq 1$ は重みであり，これが大きいほど リンクを重視したクラスタリングになり，小さいとキーワード すなわちページの内容を重要視したクラスタリングとなる．式 (6) は固有値問題には帰着できないので前章までと同様に反復法で解く．Lagrange 乗数法より式 (6) から

$$x_{1i} = \frac{z_{1i}}{\sqrt{\sum_{i=1}^{m} v_i z_{1i}^2}} \quad (7)$$

$$y_{1j} = \frac{\sum_{i=1}^{m} v_i w_{ij} x_{1i}}{\sqrt{\sum_{j=1}^{n} v_j \left(\sum_{i=1}^{m} v_i w_{ij} x_{1i} \right)^2}} \quad (8)$$

が導かれる（付録参照）．ここで

$$z_{1i} = 2\alpha \sum_{i'=1}^{m} w_{ii'} v_{i'} x_{1i'} + (1-\alpha) \sum_{j=1}^{n} w_{ij} v_j y_{1j} \quad (9)$$

である．式 (7) と (8) を反復計算して解を求める．まず x_1 と y_1 の初期値を任意に設定し，それを式 (7) の右辺に代入して x_1 を求め，それを式 (8) に代入して y_1 を求め，これらの新しい x_1 と y_1 を式 (7) に代入することを x_1 と y_1 が収束するまで繰り返す．この反復法の収束性の証明はここでは省略するが，Liapunov の安定性定理によって，初期値によらずに大域的に収束することが示せる．

以上の反復で得られる x_1 と y_1 とから $p_{1i} = x_{1i}/\max\{x_{1i}\}$，$q_{1j} = y_{1j}/\max\{y_{1j}\}$ によりページのメンバシップ p_{1i} とキーワードのメンバシップ q_{1j} が得られる．

次に第 2 クラスタ抽出では，今はページ集合をクラスタリングしているのであるから，ページ i の重みが $(1-p_{1i})v_i$ と削減され，キーワード j の重みは v_j のままとする．したがって，各ページとキーワードが第 2 クラスタに所属する割合 x_{2i} と y_{2j} は，式 (6) の v_i を $(1-p_{1i})v_i$ に変えて同じことをすれば求まる．第 3 クラスタ以降も同様であり，一般に第 k クラスタでは式 (6) の v_i を $\prod_{l=1}^{k-1}(1-p_{li})v_i$ として同じことを行えばよい．この場合も抽出されるクラスタの凝集度は順番が進むに従って単調に減少するので，この凝集度の変化に基づいてクラスタ数を決める．

5.2 実　験　例

例として 4 章の 118 個のウェブページに 28 個のキーワードを付け加えてページをクラスタリングした．式 (6) の α は 0.5 とした．ページと代表キーワードの配置を**図 8** に示す．このクラスタリングにおいても 3 章と同様にキーワード 割付けのノイズが平滑化される．すなわちあるキーワードに関連深いページであるにもかかわらず，たまたまそのキーワードが割り付けられなかったような場合でも，全体をクラスタリングすることによってそのページとキーワードとの関連性は大きくなる．今の場合この平滑化作用は 2 重に行われる．

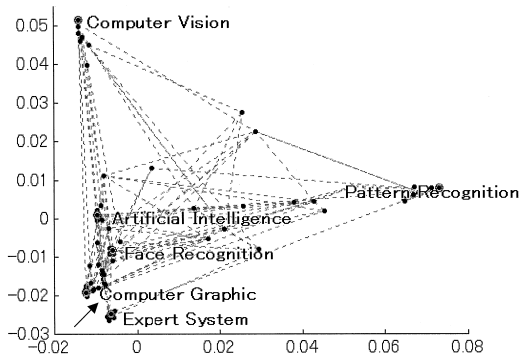


図 8 ページと代表キーワードの配置
Fig. 8 Arrangement of web pages and representative keywords.

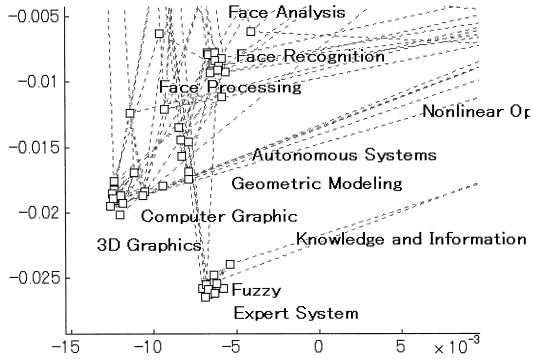


図 9 図 8 の矢印付近の拡大図
Fig. 9 Magnification of place directed by arrow in Fig. 8.

1つは3章のものであり，各ページへのキーワードの分布状況に基づいて生じ，もう1つはクラスタリングのときにリンクとキーワードの両方を用いていることにより，あるキーワードを含まないページでもリンクを通してそのキーワードと関係が生じることによる．このような高度な平滑化は LSI 法では困難である．**図 9** は図 8 の左下を拡大表示したものである．ここで得られたメンバシップに基づいて3章の終わりに書いたハンティング検索をしてみた．Artificial Intelligence, Intelligent information, Heuristic Method の3つのキーワードで OR 検索をし，s_i が大きいページを 10 個選んで表示したのが**図 10** である．z 座標は s_i の値である．どのキーワードとのコサインが最大であるかによって各ページを色と形で区別している．この検索法では，上記の平滑化作用により，クエリキーワードを含まないページでもリンクを通してそのキーワードと関係があるようなページを検索することができる．

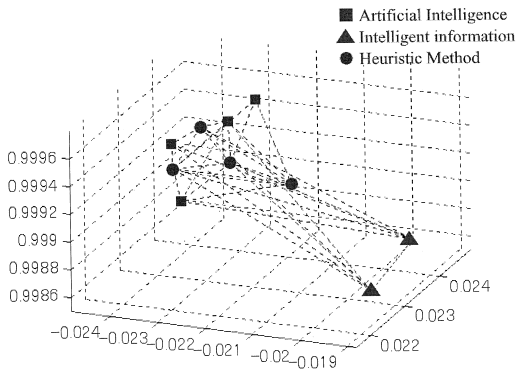


図 10 検索ページの表示
Fig. 10 Display of retrieved web pages.

6. むすび

データをファジークラスタリングし，それを用いてデータを検索したり，クラスタ構造を表示したりする方法をグラフ構造データに応用して，さらに複雑な構造のグラフに拡張した．本方法ではファジークラスタリングで得られるメンバシップをハンティング検索やウェブリンクの推薦などに利用し，またメンバシップ値に基づいてデータを数量化3類で配置表示することによりデータのクラスタ構造を視覚化し，ブラウジング検索や視覚的なリンクの推薦などに利用する．本検索法は LSI と同様にクエリとの表面的な一致性でなく潜在的な関連性で検索することができる．また数量化3類による配置法は線形であるから他の非線形な配置法に比べ，位置関係のひずみは大きいと思われるが計算の手間は少ない．ここでは比較的少数のデータの検索や表示を扱ったが，データ数が大量になると階層化などの構造化を行わなければ検索も表示も効率が落ちると思われる．ここで提案した方法を階層的クラスタリングに拡張して大量データに応用するのが今後の課題である．

参考文献

1) Florescu, D., Levy, A. and Mendelzon, A.: Database Techniques for the World-Wide Web, A Survey, *SIGMOD Record*, Vol.27, No.3, pp.59–74 (1998).
2) Guillaume, D. and Murtagh, F.: Clustering XML Documents, *Comput. Phys. Comm.*, Vol.127, No.2/3, pp.215–227 (2000).
3) Mancoridis, S., Mitchell, B.S., Rorres, C., Chen, Y. and Gansner, E.R.: Using Automatic Clustering to Produce High-Level System Organizations of Source Code, *6th Int. Workshop*

Program Understanding, pp.34–41 (Jun. 1998).
4) Cvetkovic, D.M., Doob, M. and Sachs, H.: *Spectra of Graphs*, Academic Press, New York (1980).
5) Inoue, K. and Urahama, K.: Sequential Fuzzy Cluster Extraction by a Graph Spectral Method, *Patt. Recog. Lett.*, Vol.20, No.7, pp.699–705 (1999).
6) Hotta, S., Inoue, K. and Urahama, K.: Extraction of Fuzzy Clusters from Weighted Graphs, *Proc. PAKDD-2000*, Kyoto, Terano T. (Ed.), pp.442–453, Springer-Verlag (2000).
7) 井上光平，浦浜喜一：共起関係行列に基づくファジークラスタリングとデータ検索への応用，電子情報通信学会論文誌，Vol.J83-D-II, No.3, pp.957–966 (2000).
8) 堀田政二，井上光平，浦浜喜一：ファジィクラスタ構造の可視化，映像情報メディア学会論文誌，Vol.54, No.2, pp.319–321 (2000).
9) 園川隆夫：多変量のデータ解析，朝倉書店 (1988).
10) Kohonen, T.: *Self-Organizing Maps*, Springer-Verlag, Berlin Heidelberg (1995).
11) Eades, P.: An Algorithm for Drawing General Undirected Graphs, *Cong. Num.*, Vol.42, pp.149–160 (1984).
12) Deerwester, S., Dumais, S., Furnas, G., Landauer, T. and Harshman, R.: Indexing by Latent Semantic Analysis, *J. Amer. Soc. Inf. Sci.*, Vol.41, pp.391–407 (1990).
13) Kleinberg, J.: Authoritative Sources in a Hyperlinked Einvironment, *9th ACM-SIAM Symp. Disc. Alg.*, pp.668–677 (1998).

付録　式 (2)，(4)，(5)，(7)，(8) の導出

式 (1) の Lagrange 関数は

$$L=\sum_{i=1}^{m}\sum_{j=1}^{m}v_i x_{1i} w_{ij} v_j x_{1j} -\lambda\left(\sum_{i=1}^{m} v_i x_{1i}^2 - 1\right) \tag{10}$$

となる．λ は Lagrange 乗数である．式 (1) の解は

$$\frac{1}{2}\frac{\partial L}{\partial x_{1i}}=\sum_{j=1}^{m} v_i w_{ij} v_j x_{1j} - \lambda v_i x_{1i} = 0 \tag{11}$$

$$\frac{\partial L}{\partial \lambda}=1-\sum_{i=1}^{m} v_i x_{1i}^2 = 0 \tag{12}$$

を満たす．式 (11) から

$$x_{1i}=\sum_{j=1}^{m} w_{ij} v_j x_{1j}/\lambda \tag{13}$$

となるから，これを式 (12) に代入して λ を求めると

$$\lambda=\sqrt{\sum_{i=1}^{m} v_i \left(\sum_{j=1}^{m} w_{ij} v_j x_{1j}\right)^2} \tag{14}$$

となるから，これを式 (13) に代入すると式 (2) が得られる．

式 (4)，(5) についても同様に，式 (3) の Lagrange 関数は

$$L=\sum_{i=1}^{m}\sum_{j=1}^{n} v_i x_{1i} w_{ij} v_j y_{1j} - \lambda\left(\sum_{i=1}^{m} v_i x_{1i}^2 - 1\right) -\mu\left(\sum_{j=1}^{n} v_j y_{1j}^2 - 1\right) \tag{15}$$

となる．式 (3) の解は

$$\frac{\partial L}{\partial x_{1i}}=\sum_{j=1}^{n} v_i w_{ij} v_j y_{1j} - 2\lambda v_i x_{1i} = 0 \tag{16}$$

$$\frac{\partial L}{\partial y_{1j}}=\sum_{i=1}^{m} v_i w_{ij} v_j x_{1i} - 2\mu v_j y_{1j} = 0 \tag{17}$$

$$\frac{\partial L}{\partial \lambda}=1-\sum_{i=1}^{m} v_i x_{1i}^2 = 0 \tag{18}$$

$$\frac{\partial L}{\partial \mu}=1-\sum_{j=1}^{n} v_j y_{1j}^2 = 0 \tag{19}$$

を満たす．上で式 (2) を導いたのと同様にして，式 (16) から

$$x_{1i}=\sum_{j=1}^{n} w_{ij} v_j y_{1j}/2\lambda \tag{20}$$

となるから，これを式 (18) に代入すると λ が求まるのでそれを再び式 (20) に代入すると式 (4) が得られる．同様にして式 (17) と式 (19) とから μ を消去すると式 (5) が得られる．式 (7)，(8) も同様にして式 (6) から導かれる．

(平成 12 年 6 月 20 日受付)
(平成 12 年 9 月 27 日採録)

(担当編集委員　吉川 正俊)

堀田 政二

1999 年九州芸術工科大学大学院博士前期課程修了．現在，同大学院博士後期課程在学中．画像情報処理とその応用に関する研究に従事．

井上 光平（正会員）

2000 年九州芸術工科大学大学院博士後期課程修了．現在，九州芸術工科大学助手．パターン認識，画像処理に関する研究に従事．

浦浜 喜一

1980 年九州大学大学院工学研究科博士後期課程修了．現在，九州芸術工科大学教授．パターン認識，画像情報処理に関する研究に従事．