クラスタリングに基づく情報の検索と視覚化

堀田, 政二

https://doi.org/10.15017/1398256

第6章

グラフ構造データの検索と視覚化

6.1 まえがき

クエリに類似したデータを検索するハンティング検索において，例えばキーワードに基づくデータ検索では，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はなく，データ全体に渡る大域的なキーワード付与分布に基づいて検索を行うことが重要である [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 Latent Semantic Indexing(LSI)[9] などの手法がある．しかし，クラスタリングに基づく検索ではこれまで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が失われるので，クラスタレベルでの粗い検索しか行えなかった．

一方，グラフの視覚化については多くの研究がされてきている [12] が，ほとんどの手法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．例えば，ウェブ [35] やプログラム [36] を有向グラフとして表し，クラスタリングを行ってデータ構造を視覚化する研究も行われているが，これらでもクラスタリングと視覚化とは結合されておらず，グラフを表示するときにはグラフ自動描画法が使われている．これらの研究でクラスタリングと視覚化とが結合していない理由は，クラスタリングがハードであることである．ハードクラスタリングでは各データの各クラスタへのメンバシップは0か1であるので，クラスタの中でデータを並べ替えることは無意味である．したがって，クラスタ内部のデータは列記するくらいしかなく，クラスタリングに基づく多くのブラウジング検索システムではそのような表示しか行われておらず，視覚的にデータの構造は把握し難い．

そこで本章では，グラフ構造データをファジークラスタリングし，各データのメンバシップに基づいてハンティング検索を行う方法と，数量化3類でグラフ構造デー

タを視覚化する方法を提案する [25, 26, 27]. ファジークラスタリングの方法としては，データ構造が無向グラフや2部無向グラフあるいは有向グラフである場合にはグラフスペクトル法の一種である第3章の方法を，混成グラフの場合は反復法である第4章の方法を用いる．まずはじめに，提案手法の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は，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．

6.2 ファジークラスタリングを利用したハンティング検索

まずはじめに，ファジークラスタリングを利用したハンティング検索について，LSI法 (Latent Semantic Indexing method) と比較しながら説明する．例として表 6.1 のようなデータ行列が与えられたとする．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．通常のハンティング検索では，ユーザがキーワード，あるいは文献自身をクエリとして入力し，クエリと文献との類似度を計算した後に，類似度が大きい文献から順番に出力する方法が一般的である．類似度の尺度としてはコサイン値（余弦尺度）が利用される．このコサイン値による順位付けはベクトル空間

表 6.1: 共起関係行列の例

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

法 [8] にも使われている方法である．例えば，文献が m 個，キーワードが n 個あるとする．クエリを $\boldsymbol{q} = [q_1, ...q_n]^T$ とし，文献 j を $\boldsymbol{d}_j = [d_{j1}, ... d_{jn}]^T$ とすれば，クエリと文献とのコサイン値は

$$\cos\theta_j = \frac{\boldsymbol{d}_j^T \boldsymbol{q}}{\|\boldsymbol{d}_j\|\|\boldsymbol{q}\|} \tag{6.1}$$

により求められる．この $\cos\theta_j$ が大きい順に文献を出力する．ここで $\boldsymbol{d}_j^T \boldsymbol{q}$ は内積 $\sum_{i=1}^{n} d_{ji} q_i$ であり，$\| \boldsymbol{d}_j \|$ はユークリッドノルム $\sqrt{\sum_{i=1}^{n} d_{ji}^2}$ である．例えば，キーワード 10 をクエリとして入力し，表 6.1 をそのまま使って各文献とのコサイン値を求めると図 6.1 のようになる．コサイン値が大きな値をとる文献は文献 6. 7. 8 であり，その他の文献はキーワード 10 を含まないのでコサイン値はすべて 0 となる．

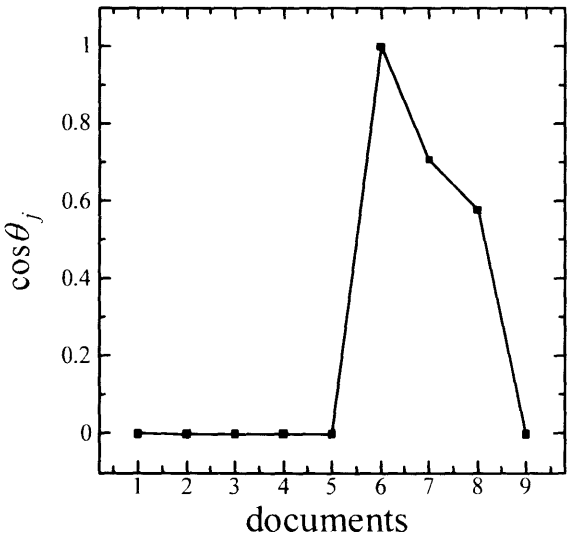


図 6.1: 各文献のコサイン値

しかし，表 6.1 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，これらのキーワードを含んでいる文献 6. 7. 8 等と内容的には類似していると推測できるからである．そこで，このような潜在的な関係に基づいて検索を行う潜在意味解析法 (Latent Semantic Indexing method. LSI 法) が提案された [9]. LSI 法は，表 6.1 のような共起関係行列を特異値分解し，少数の次元で元の共起関係行列を近似して，ノイズを平滑化すると同時に類似度の計算量を削減する方法である．LSI 法における類似度も式 (6.1) の余弦尺度で求める．LSI 法でキーワード 10 をクエリとしたときの各文献のコサイン値を図 6.2 に示す．図を見ると，表 6.1 をその

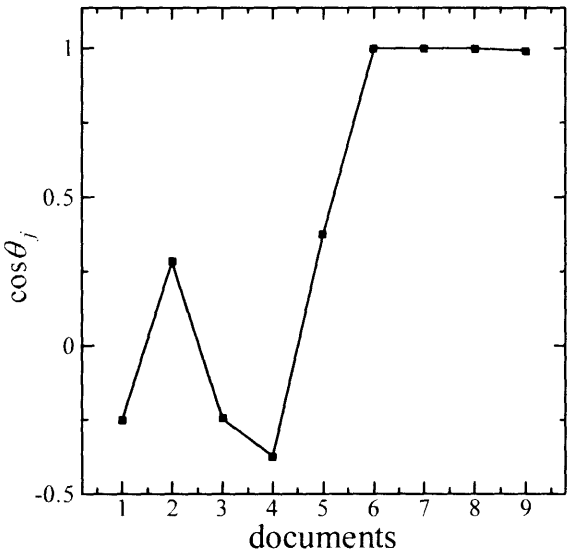


図 6.2: LSI 法による各文献のコサイン値

まま使った検索とは異なり，文献 6, 7, 8 に加えて文献 9 のコサイン値も大きくなっている．LSI 法ではこのように，潜在的な関連に基づいた検索を行うことができる．

本章で提案するハンティング検索法も，LSI 法と同様に潜在的な関連を反映した検索結果を得ることができる．大きな相違点は，LSI 法では特異値分解を利用して大域的な情報抽出を行うが，提案手法はクラスタリングを利用して潜在的な関連を抽出する点である．すなわち，提案手法では共起関係行列に基づいて予め文献をクラスタリングし，文献とキーワードの各クラスタへのメンバシップを求めておく．クラスタの数を N としておく．クエリとしてキーワードが与えられた場合，そのキーワードの各クラスタへのメンバシップ q_i $(i = 1, ..., N)$ のベクトル $\boldsymbol{q} = [q_1, ..., q_N]^T$ と，文献 j の各クラスタへのメンバシップ d_{ji} のベクトル $\boldsymbol{d}_j = [d_{j1}, ..., d_{jN}]^T$ とのコサイン値を式 (6.1) で求めて，コサイン値が大きい順に文献を出力する．この方法はクエリが所属するクラスタと同じクラスタに所属する文献を選ぶという選択法をファジー化したものであるともいえる．この順位付けではクエリに最もよく似たメンバシップ分布を持つ文献が 1 位になる．文献自身がクエリとして与えられる場合も全く同じで，その文献の各クラスタへのメンバシップのベクトルと，他の文献のメンバシップのベクトルとのコサイン値が大きな文献から順に選び出せばよい．

LSI 法と比較するために実験を行った．はじめに，表 6.1 に基づいて第 3 章 3.4.1 節の方法でクラスタリングを行った．凝集度の変化を図 6.3 に示す．横軸はクラスタの番号である．この図では 4 個めから 5 個めへの変化は確実に小さいが 3 個めから 4 個めの変化も小さいと見るかどうかは曖昧である．ここでは 3 個めからも小さいと見る

ことにすると主要なクラスタの数は2個となり，文献[9]とも一致する．図6.4が本方法で得られたクラスタへのメンバシップである．図6.4(a).(b)が第1クラスタへのメンバシップ，(c).(d)が第2クラスタへのメンバシップである．(a).(c)が文献(横軸は文献の番号)，(b).(d)がキーワードのメンバシップである．図6.4(a).(c)から抽出されたクラスタは文献1から5と文献6から9であることがわかる．また，図6.4(b).(d)からわかるようにキーワードも1から8と10から12に分かれている．キーワード9はどちらのクラスタにも同程度所属する．これは表6.1からも妥当な結果である．

このクラスタリングに基づいて，キーワード10をクエリとして各文献のコサイン値を求めると図6.5のようになった．キーワード10と同じクラスタに所属する文献6から9が検索されていることがわかる．特に文献9は元々キーワード10を含んではいなかったが，クラスタリングの結果キーワード10に対して高いコサイン値を示すようになっている．これはLSI法などと同じ効果である．

同じような効果は文献をクエリとする場合でも現れることを検証する．図6.6は文献9をクエリとしたときの各文献のコサイン値である．文献9と同じクラスタに所属する文献(6から9)が検索されることがわかる．特に文献2と文献6のコサイン値に注意されたい．表6.1からわかるように文献6は文献9と共通するキーワードを一つも持たない．それに対し文献2は文献9と共通のキーワードを一つ持つ．したがって，文献9をクエリとして表6.1から直接検索を行うと文献2の方が文献6よりもコサイン値が大きくなる．図6.6はこれとは逆である．これは文献6は文献7や8を通して文献9と関連を持つということをクラスタリングにより抽出できたからである．以上のように，本方法は間接的な関連性も含めた大域的な情報を反映した検索が行える．

なお，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と文献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が大きい文献から順に選ぶ．なお統合法としてこの他にも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，後に示す6.6.1節での実験ではminとmaxの方が良好な結果が得られた．

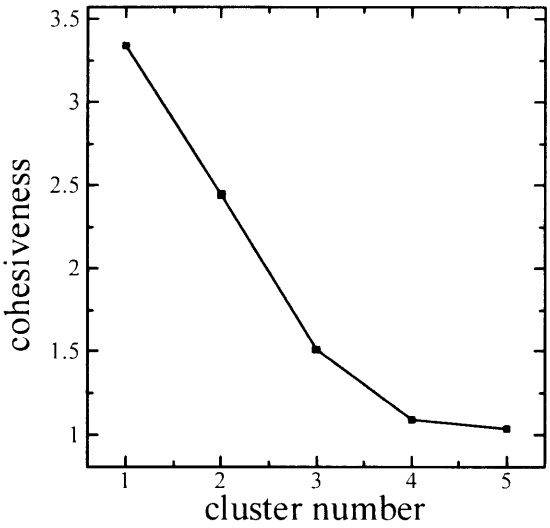


図 6.3: 凝集度の変化

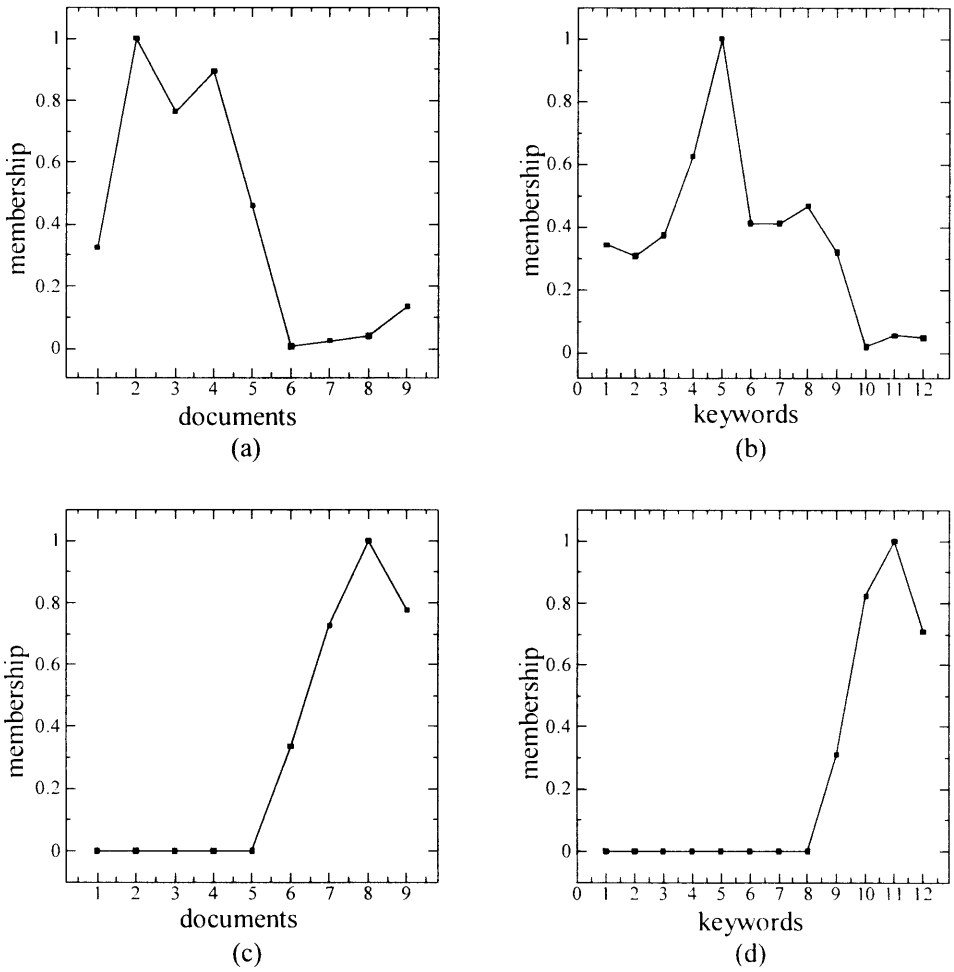


図 6.4: 文献とキーワードのクラスタへのメンバシップ

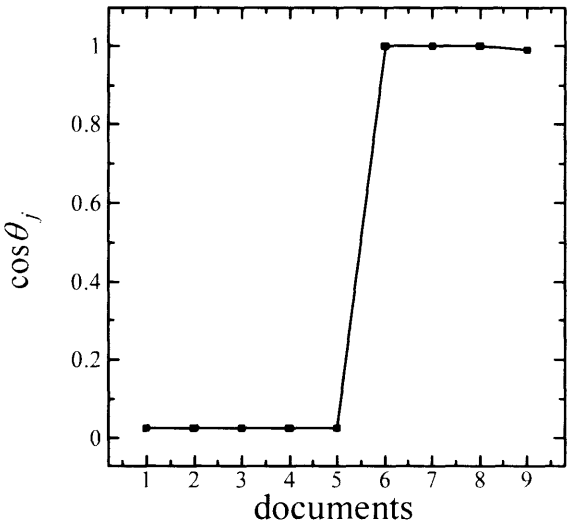


図 6.5: キーワード 10 をクエリとした場合の各文献のコサイン値

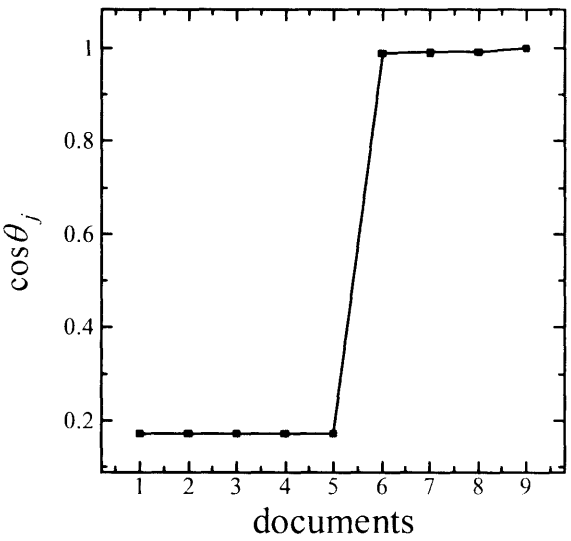


図 6.6: 文献 9 をクエリとした場合の各文献のコサイン値

6.3 ファジークラスタリングを利用したブラウジング検索

ここでは，ファジークラスタリングを利用してブラウジング検索を行う方法を示す．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．そこで，数量化3類によってそのようなデータ配置を求めることにする．数量化3類を用いる利点とデータを配置する具体的な方法は既に第5章で述べているが，ここでもう一度概説しておく．

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，$m \leq n$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$\boldsymbol{u}_2, \boldsymbol{u}_3$を求めて$\boldsymbol{x}=F^{-\frac{1}{2}}\boldsymbol{u}_2, \boldsymbol{y}=F^{-\frac{1}{2}}\boldsymbol{u}_3$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$\boldsymbol{v}_2, \boldsymbol{v}_3$と固有値$\lambda_2, \lambda_3$を求めて$\boldsymbol{x}=F^{-1}DG^{-\frac{1}{2}}\boldsymbol{v}_2/\sqrt{\lambda_2}, \boldsymbol{y}=F^{-1}DG^{-\frac{1}{2}}\boldsymbol{v}_3/\sqrt{\lambda_3}$とすれば$(x_i, y_i)$が第$i$個体の2次元座標となる．ファジークラスタリングの結果を表示するには個体をデータ，項目をクラスタに置き換えd_{ij}をクラスタへのメンバシップ値とすればよい．この場合には数量化4類よりも良好なデータ配置が得られ，計算量も少なくてすむ．

第5章では主にデータを2次元平面に配置する方法を述べたが，ここでは3次元空間にデータを配置表示することを考える．ここで第3軸（一般にz座標）として何を選ぶのか考えた場合，主に次の2つが候補として挙げられる；1) 数量化3類の第4固有ベクトルを第3軸とする．2) データの各クラスタへのメンバシップ値を第3軸とする．このうち1) はデータ間の関係が既にxy座標である程度与えられているので，ユーザに新たな情報を与えるという観点からみれば有効な手段ではない．一方，2) では各データの主要度も視覚的に把握することができるので有用な方法であるといえる．そこで本章では2) の方法を使用する．また，ハンティング検索の結果も3次元表示することにする．すなわち，クエリとデータとのスコアを第3軸として表示することにより，クエリとの類似性を視覚的に把握できるようにする．

なお，上記の表示法はLSI法でも可能である[37]．しかし，混成グラフにはLSI法を適用することができないので，複雑なグラフ構造データの視覚化は不可能である．一方，本章の提案手法はデータのメンバシップ値に基づく配置方法なので，複雑なグラフ構造データでも視覚化は可能である．

6.4 2部無向グラフの視覚化と検索

ここでは2部無向グラフで表されるデータを使った例として，画像とキーワードを同時に表示して対象を絞り込むブラウジング検索法と，視覚的協調フィルタリングによる映画の推薦への応用例を示す．

6.4.1 画像とキーワードの例

ここではキーワードによる画像の検索について実験を行った結果を示す．まず，198枚の画像に54個のキーワードを割り振ったデータを作成した．画像とキーワードとの対応関係を2部無向グラフで表し，3.4.1節の方法で画像をクラスタリングした．凝集度の変化は図6.7のようになり，これからクラスタの数は5とした．抽出された5個のクラスタの代表画像と代表キーワードを図6.8に示す．クラスタリング結果の表示は，ハードクラスタリングの場合は通常図6.8のように各クラスタの代表データを羅列するだけである．これではクラスタ同士の類似度や各データ間の類似関係などはわからない．そこで各クラスタへのメンバシップに基づいて，198枚の画像と54個のキーワードを数量化3類により2次元平面に同時配置したものを図6.9に示す．画像は白い四角，キーワードは黒点であり，代表画像と代表キーワードを大きく表示している．辺は画像とキーワードの対応関係を示している．このようにすると

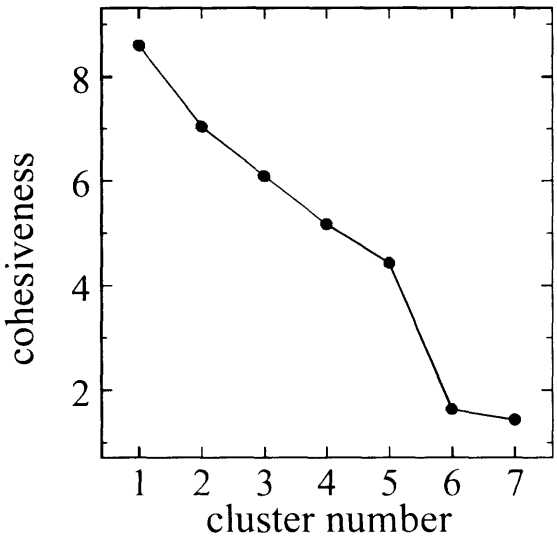


図6.7: 凝集度の変化

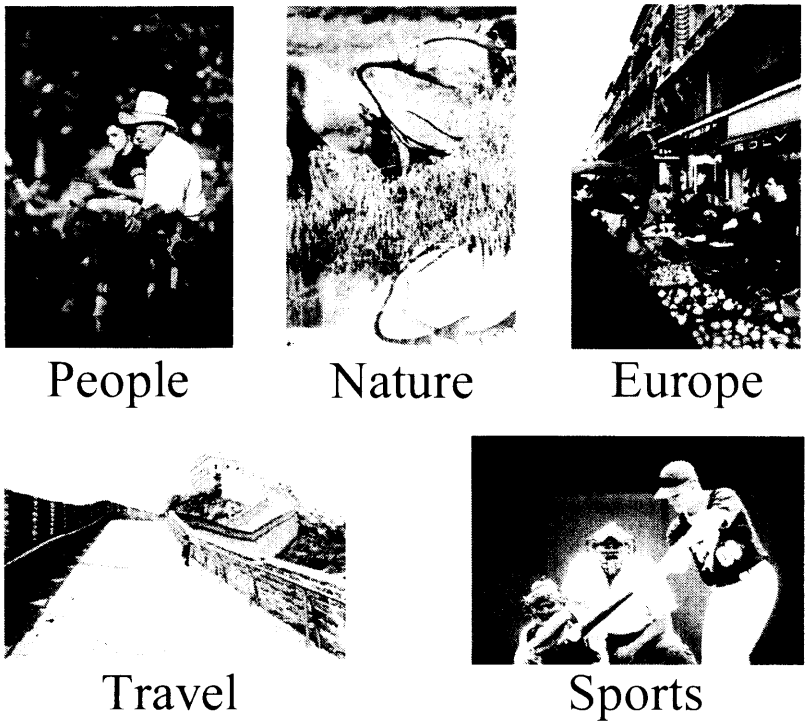


図 6.8: クラスタの代表画像と代表キーワード

図 6.8 のように代表データを単に羅列するよりもクラスタどうしの関係がつかめ，また各データのクラスタ内での位置もある程度わかる．

図 6.10 は図 6.9 の中央付近の People のクラスタを拡大表示したものである．見づらくなるので辺は省略した．このように画像とキーワードを同じ画面に表示すれば両者の対応関係が分かり，キーワードでサーチして画像を検索するといったことができる．このようなデータ表示法は視認性を考慮したグラフ描画法 [12] とは表示目的が異なるが，クラスタ構造に注目したグラフの描画法の 1 種であるといえる．

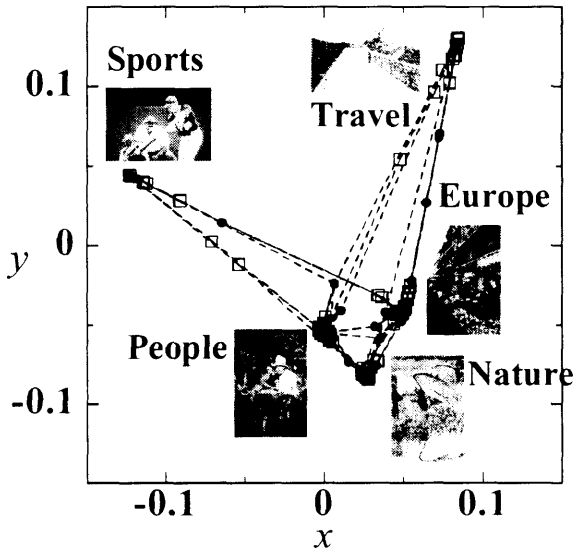


図 6.9: 画像とキーワードのクラスタリング結果の表示

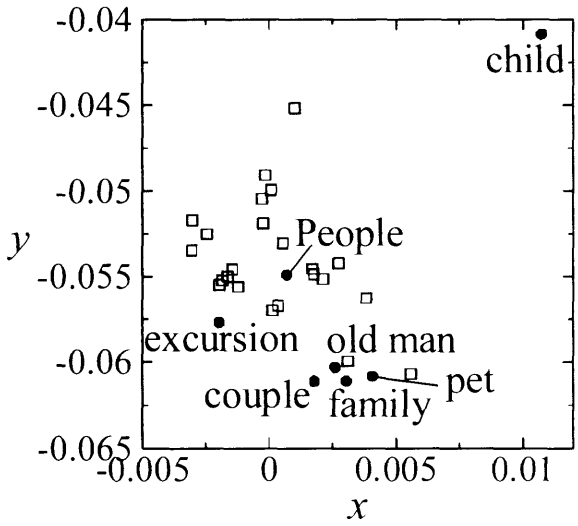


図 6.10: クラスタの拡大表示

6.4.2 協調フィルタリングへの応用例

ここでは視覚化の応用例として協調フィルタリングへの利用を試みる．協調フィルタリングとは，各人の嗜好に基づいて製品を推薦する際に，嗜好が似ている他の人のデータを利用して推薦度を求める手法である．相関法や機械学習など数値的なアルゴリズムがこれまで開発されているが，スプリングモデルなどの配置法を使った視覚的な手法も試みられている [38]．例として，協調フィルタリングの検証用データとしてよく用いられるEachMovie[1]について実験してみた．EachMovieは72916人の1628の映画に対する6段階評価 (0.0,0.2,0.4,0.6,0.8,1.0) のデータである．すなわち，データは 72916×1628 の行列であるが，各人が見た映画はそれほど多くないので行列のほとんどの要素は空である．そのようなまだ見ていない映画について各人の嗜好度を推定して推薦するのが協調フィルタリングの目的である．ここではEachMovieデータのうち最初の3000人の500本の映画に関するデータを用いた．人と映画を点として，評価値が与えられている人と映画とを辺で結び，辺の重みを評価値として人を3.4.1節の方法でクラスタリングした．3個のクラスタを抽出して，そのメンバシップに基づいて人と映画とを2次元平面へ配置した結果の1部分を拡大表示して図6.11に示す．この図で自分の周囲を見れば好みに合う映画が近くにあり，また嗜好が似ている人も近くに位置する．したがって，自分の近くにある映画でまだ見ていないものは今後見るべき映画の候補となり得る．この配置の有効性について検証してみた．各人について評価値が高い映画が自分の近くに来ていればこの配置は有効であるといえる．評価値のデータがない映画は検証には使えない．そこで，各人の近くにある10個の映画のうち，その人が評価値を付けている映画について評価値の平均値を求めたところ0.7となり，まずまずの値が得られた．したがって，各人の近くにはその人が好きな映画が配置されており，この表示により視覚的な協調フィルタリングを行うことができるものと思われる．

なお，このデータのクラスタはかなり分散が大きく，したがって，本方法で得られる配置はデータがかなり一様に分布し，スプリングモデルで得られる配置とよく似た結果が得られる．ただし，計算時間は本方法がスプリングモデルよりも約千倍速い．これはスプリングモデルが全データの座標を逐次に動かして行くのに対し，本方法ではデータの数に関らずクラスタ数 (この例では 3×3) の行列の固有値を求めるだけでよいからである．上記の近傍10個の映画の評価値の平均はスプリングモデルでは0.75となり，これに比べて本方法はそれほど遜色ない配置であるといえよう．また，

[1] http://www.research.compaq.com/SRC/eachmovie/

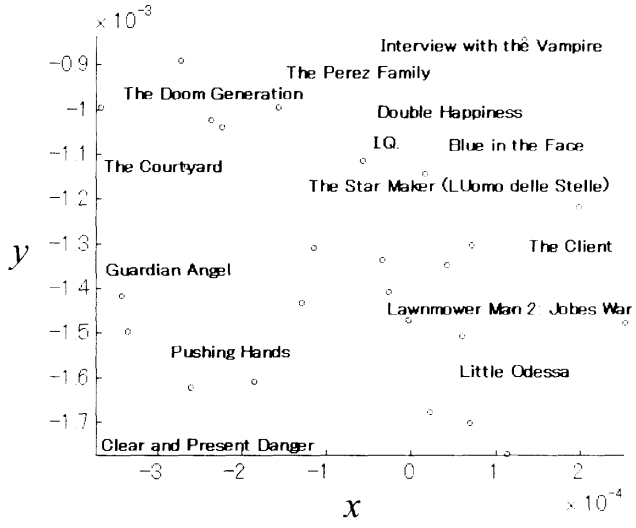


図 6.11: 人と映画の配置の拡大図

スプリングモデルではデータの配置だけしか得られないが，本方法では各データのメンバシップも得られるので，各グループを代表する人や映画などを知ることもできる．

6.5 有向グラフの視覚化と検索

有向グラフの視覚化と検索の例として，ウェブページのリンク関係を利用した場合を示す．"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3.2節の方法でクラスタを抽出した．図6.12に抽出されたクラスタの凝集度の変化を示す．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個であることがわかる．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6.13(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．なお各クラスタの代表画像の位置は必ずしもクラスタの中央付近になるとは限らない．次に有向グラフとして3.3節の方法で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6.13(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6.13(c)に示す．図6.13(a).(b).(c)とも，だいたい7個

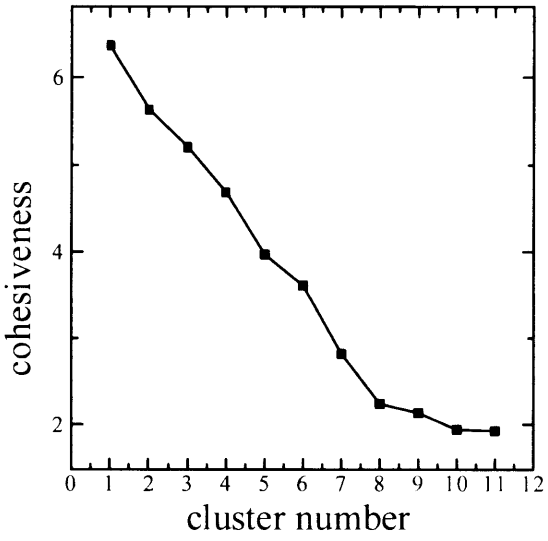
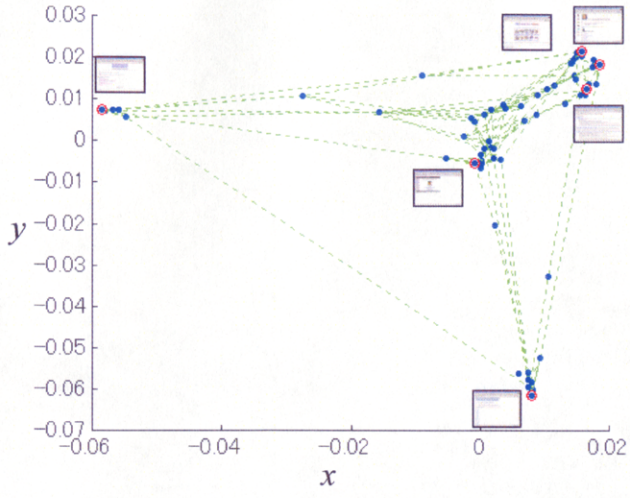


図 6.12: 凝集度の変化

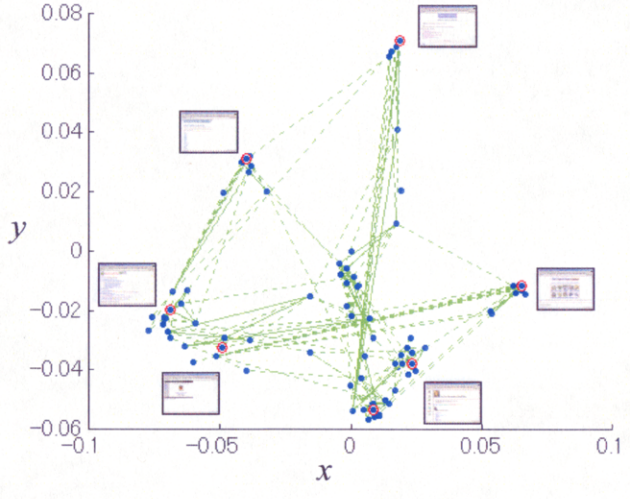
のクラスタからなることがわかる．図 6.13(a) ではリンクの向きによらず多数のリンクに接続するページのメンバシップ値が大きい．また，図 6.13(b) ではクラスタ内のページの多くにリンクを張っているページがクラスタの代表となる．このようなページを Kleinberg ら [39] はハブと呼んでいる．一方，図 6.13(c) では多くのページからリンクを張られているページがクラスタの代表となる．そのようなページはオーソリティと呼ばれる [39]．これらハブやオーソリティは各ページのリンク数をカウントするだけでは検出できない．他のページを介した間接的なリンクも考慮しないといけないからである．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．図 6.14 は xy 座標を図 6.13 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 6.14(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 6.14(b) では赤の鎖線が多く，図 6.14(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．例えば図 6.15 は図 6.13(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関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 6.2 節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言える．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関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．

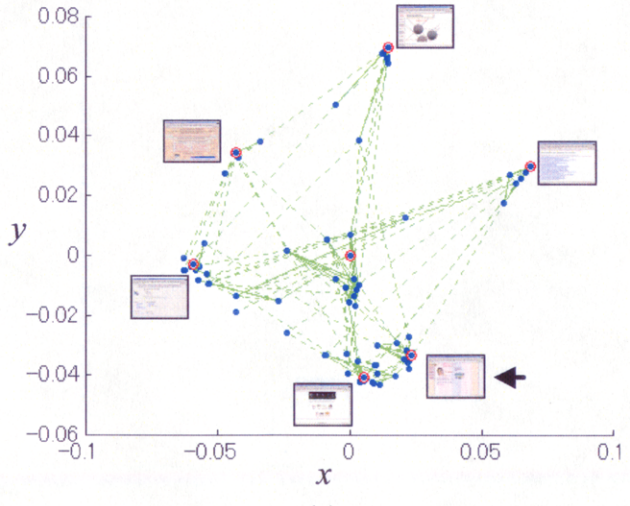
なお，比較のために数量化 4 類，自己組織化ネット (SOM)，スプリングモデルでも実験してみた．SOM ではデータが特徴ベクトルとして与えられる必要があるのでメンバシップを特徴ベクトルとして SOM を適用した．計算時間は本方法 (数量化 3 類) が 0.01 秒，数量化 4 類も 0.01 秒，SOM が 2.58 秒，スプリングモデルが 18.1 秒であった．本方法の配置結果は図 6.13(a) であるが，数量化 4 類ではクラスタの相関関係に歪が生じる．図 6.16 に SOM による配置結果を示す．見やすいようにリンクは省略してある．○や△等はクラスタの違い（クラスタは 7 個）を表す．図 6.16 の配置は図 6.13(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の分布はつかみにくい．スプリングモデルでも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．



(a)



(b)



(c)

図 6.13: 数量化 3 類によるページの配置

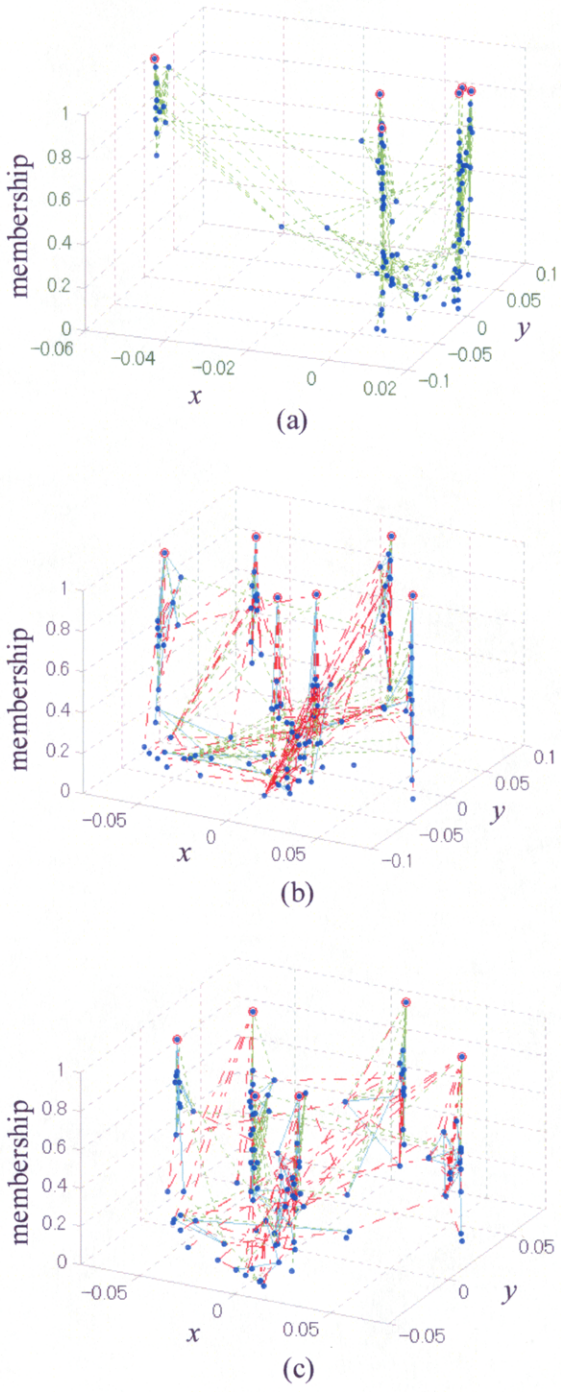


図 6.14: メンバシップを z 座標とした表示

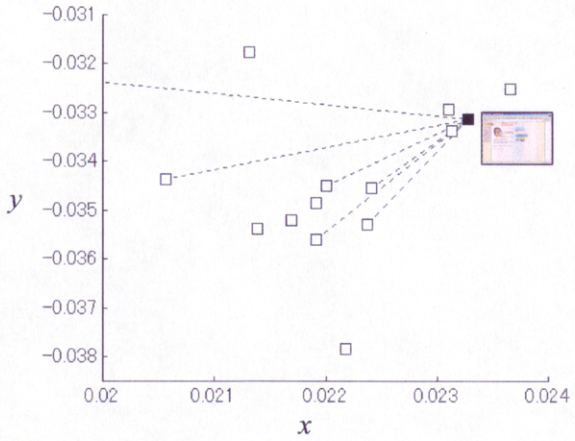


図 6.15: 図 6.13(c) の右下付近のクラスタの拡大図

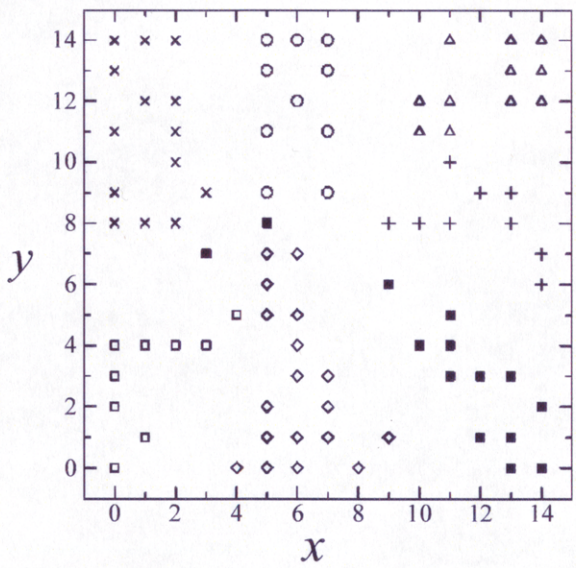


図 6.16: SOM によるウェブページの配置

6.6 混成グラフの視覚化と検索

ここでは，データが混成グラフで与えられる場合の視覚化と検索の例を示す．はじめに，ウェブページの検索と視覚化例を示す．前節の有向グラフの場合ではリンク情報のみを扱ったが，更に各ページのキーワードも与えられるとする．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，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．このデータから第4章の反復法によりファジークラスタを抽出し，ページとキーワードを同時に配置して検索を行う．次に，印象によるソファとカーペットの検索と，製品のコーディネートに視覚化を利用する感性検索への応用例をあげる．

6.6.1 ウェブの検索

例として前節の118個のウェブページに28個のキーワードを付け加えて4.5節の方法でページをクラスタリングした．式(4.17)のαは0.5とした．ページと代表キーワードの配置を図6.17に示す．このクラスタリングにおいても6.2節と同様にキーワード割り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関らず，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．1つは6.2節のものであり，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6.18は図6.17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6.2節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6.19である．z座標はs_iの値である．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．

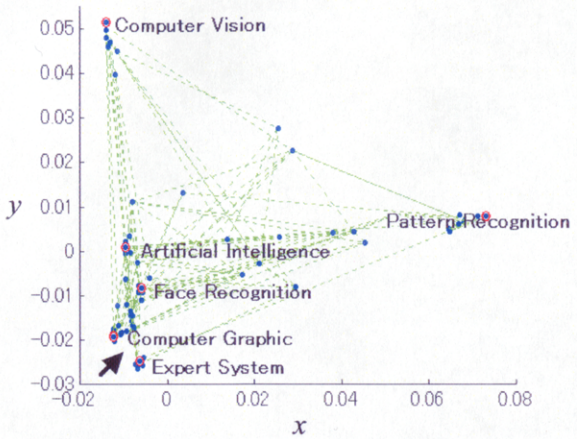


図 6.17: ページと代表キーワードの配置

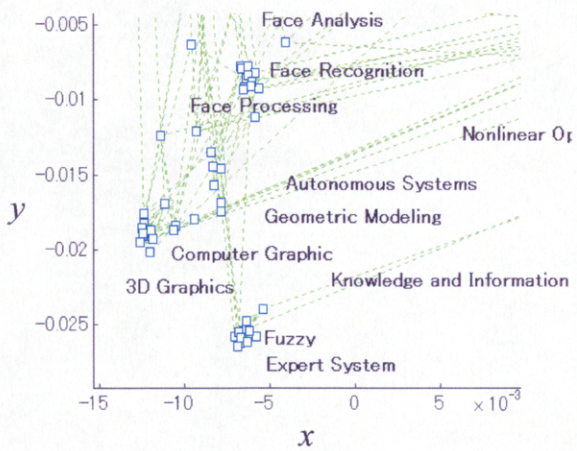


図 6.18: 図 6.17 の矢印付近の拡大図

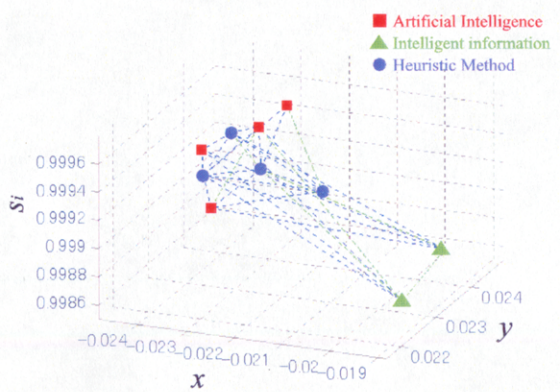


図 6.19: 検索ページの表示

6.6.2 感性検索への応用例

ここではソファとカーペットに印象 (明るい，すがすがしい，古風な等) が付いており，ソファとカーペット，および印象を同じ空間に配置することにより感性検索を行う例を示す．この場合，データ構造は2部無向グラフが2つ繋がった3部無向グラフとなる．そこで，まずはじめに，このグラフ構造でのファジークラスタ抽出法を示す．l 個のソファと m 個のカーペットがあり，それらに n 個の印象の評価値が付いているとする．第 i ソファの第 k 印象の値 (0.1.2.3 の4段階) を w_{ik}，第 j カーペットの第 k 印象の値を w_{jk} とする．簡単のためデータの重みはすべて1とする．第1クラスタに所属する割合を (簡単のため添え字の1は略して)，ソファ i が x_i，カーペット j が y_j，印象 k が z_k とすると

$$\begin{aligned} \max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}} \quad & \sum_{i=1}^{l}\sum_{k=1}^{n} x_i w_{ik} z_k + \sum_{j=1}^{m}\sum_{k=1}^{n} y_j w_{jk} z_k \\ \text{subj.to} \quad & \sum_{i=1}^{l} x_i^2 = 1,\ \sum_{j=1}^{m} y_j^2 = 1,\ \sum_{k=1}^{n} z_k^2 = 1 \end{aligned} \tag{6.2}$$

で第1クラスタを求める．Lagrange 乗数法より式 (6.2) から

$$x_i = \frac{\sum_{k=1}^{n} w_{ik} z_k}{\sqrt{\sum_{i=1}^{l}(\sum_{k=1}^{n} w_{ik} z_k)^2}} \quad (i = 1, \ldots, l) \tag{6.3}$$

$$y_j = \frac{\sum_{k=1}^{n} w_{jk} z_k}{\sqrt{\sum_{j=1}^{m}(\sum_{k=1}^{n} w_{jk} z_k)^2}} \quad (j = 1, \ldots, m) \tag{6.4}$$

$$z_k = \frac{\sum_{i=1}^{l} w_{ik} x_i + \sum_{j=1}^{m} w_{jk} y_j}{\sqrt{\sum_{k=1}^{n}(\sum_{i=1}^{l} w_{ik} x_i + \sum_{j=1}^{m} w_{jk} y_j)^2}} \quad (k = 1, \ldots, n) \tag{6.5}$$

が導かれる．これら3つの式を反復計算して解を求める．まず $\boldsymbol{z} = [z_1, \ldots, z_n]^T$ を任意に初期設定し，それを式 (6.3) と式 (6.4) の右辺に代入して $\boldsymbol{x}$ と $\boldsymbol{y}$ を求め，それらを式 (6.5) に代入して $\boldsymbol{z}$ を求め，この新しい $\boldsymbol{x}$ を式 (6.3) と式 (6.4) に代入するというのを $\boldsymbol{x}$，$\boldsymbol{y}$，$\boldsymbol{z}$ が収束するまで繰り返す．得られた $\boldsymbol{x}$，$\boldsymbol{y}$，$\boldsymbol{z}$ から $p_i = x_i/\max\{x_i\}$, $q_j = y_j/\max\{y_j\}$, $r_j = z_k/\max\{z_k\}$ によりソファのメンバシップ p_i，カーペットのメンバシップ q_j と印象のメンバシップ r_k が得られる．

第2クラスタ抽出では，ソファとカーペットとをクラスタリングするものとすれば，ソファiとカーペットjの重みをそれぞれ$1-p_i$と$1-q_j$に削減し，印象の重みは1のままとして同様なことを行う．第3クラスタ以降も同様である．

例として36個のソファ，36個のカーペット，28個の印象語から3個のクラスタを抽出し，これらを配置した結果の1部分を拡大表示して図6.20に示す．このような表示により，ソファとカーペットの印象による感性検索やコーディネートなどが行える．この例における本方法の配置の有効性を検証してみた．各印象語について近くにある5個のソファやカーペットの印象の評価値の平均を求めたところ2.14となった．スプリングモデルでは2.4であり，これと比べて本方法はそれほど遜色ない配置であるといえる．

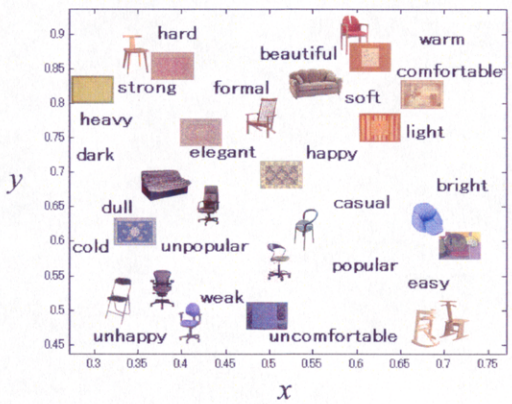


図 6.20: ソファ，カーペットと印象語の配置

6.7 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する方法をグラフ構造データに応用した．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した．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．