類似検索における特徴ベクトルのインデックスおよび関連の探索に関する一手法

姚 左軍[†] 濱田 喬[††]

特徴ベクトルを用いた画像や全文などのデータベースの類似検索において，検索の効率に及ぼす特徴ベクトルのインデックス方法は1つの大きな課題となっている．本稿では，データ間に存在する類似性に基づいた特徴ベクトルの分類法である「回帰的クラスタリング」と，この分類法で構築された特徴ベクトルの木構造インデックスに適した探索手法である「排除探索法」を提案する．本提案により特徴ベクトルを用いた画像，全文データベースの類似検索を効率的に行うことができる．すなわち，通常，検索の呼び出し率が10%以下の場合，線形検索手法による類似検索と比べ，本提案による類似検索のコストは3分の1程度にとどまることが，英文フルテキストから抽出した分野情報に基づく特徴ベクトルを用いた評価実験により確認された．

A Method for Indexing Feature Vectors and Its Related Searching Approach

ZUOJUN YAO[†] and TAKASHI HAMADA[††]

For the similarity retrieval based on feature vectors, how to construct an index of feature vectors to improve the retrieval efficiency has become an important topic. In this article we introduce a newly developed index method of feature vectors, which is composed of "recurrence clustering" and "removal search strategy." Recurrence clustering is a classification method used to construct a tree-like index of feature vectors based on similarities between feature vectors, and removal search strategy is a method tailor-developed to suit for searching the index structure constructed by recurrence clustering. The similarity retrieval based on feature vectors can be efficiently improved by this new approach. That is, the retrieval cost of this approach is less than that of linear associative retrieval strategy when the recall ratio is less than 10%. The effectiveness of the approach was confirmed in relative experiments using feature vectors extracted from full-text.

1. まえがき

画像や全文などのデータベースにおける類似検索に対して，元の生データから特徴情報を抽出し，ある表現手法，たとえば，キーワード[14]，シグネチャファイル[15]，特徴ベクトル[8]などを用いてデータの特徴を表す手法が利用されている．それらの手法を考察してみると，キーワードの表現手法は人間の感性イメージに依存するため，実際に使用するときにキーワードを決める方法は人により異なり，一般的な決定基準を求めるのは困難である．シグネチャファイルに関して，その基本要素であるシグネチャは個々の生データの特徴属性から生成される固定長のビット列であり，ビット論理演算を用いて効率的に検索処理を行うことができる．また，B^+tree データ構造を用いたシグネチャデータのインデックスを構築する案[3]もある．しかし，この手法は検索サンプルとの類似度合いによる一定範囲内のオブジェクトを探すというような類似検索を実現しにくい問題点がある．一方，特徴ベクトルの表現手法では，生データの各特徴属性をそれぞれ1つの実数で表現し，個々の生データは1つの実数ベクトルで表される．様々な特徴ベクトルの類似性を計る手法が存在し，上述の問題を簡単に解決できる．画像や全文などのデータから特徴ベクトルを生成する研究がさかんに行われ，多くの研究成果が報告されている[1),2)]．本稿では，特徴ベクトルを用いた類似検索を効率的に行うため，検索サンプルに依存しない，特徴ベクトルの間の類似性に基づく特徴ベクトルのインデックスを構築する1つの手法を提案する．

本提案は，画像やテキストの特徴の間に存在する類

† セコム情報システム株式会社セコム SC センター
SECOM Information System, SECOM SC Center
†† 学術情報センター
National Center for Science Information System

似性を基に特徴ベクトルの木構造インデックスを構成する手法である「回帰的クラスタリング」と，このようなインデックス構造に適した探索手法である「排除探索」からなる．クラスタリングは一種のデータ分類手法としてよく知られ，パタン認識などの研究[16),18)]でよく使われている．Nearest Neighbor[5)]，K-平均[17)]など多くのクラスタリングの手法が開発され，画像[6)]とドキュメント[10)]の特徴ベクトルの分類で応用されている．特に，Nearest Neighbor 法を用いた特徴ベクトルのインデックスの構成方法である階層的クラスタリングは実際のドキュメントの類似検索で使われている[11)]．

階層的クラスタリングと同様に，回帰的クラスタリングは特徴ベクトル間の類似性に基づいてデータを分類する際に木構造のインデックスを構成する．しかし，bottom-up 処理形式で特徴ベクトルの二分木構造のインデックスを構成する階層的クラスタリング[12)]と違い，回帰的クラスタリングは top-down の処理方式を用いて，cluster-tree と称する木構造インデックスを構成し，かつ，cluster-tree の分岐数をアプリケーションに応じて自由に設定できる．しかし，このようなインデックスは特徴ベクトル間の類似性に基づいたデータ分類により構成したものであるため，従来の木構造インデックスの探索手法はそれらに対して適用できなくなる．すなわち，検索処理中一部検索対象が漏れる可能性が存在する[11)]という問題が出てくる．本提案では，この問題を解決するため，新たな探索手法を再帰的クラスタリングとともに開発した．

2. 特徴ベクトルの分類

2.1 分類アルゴリズム

特徴ベクトルの類似性計算に関しては様々な方法[7),8)]がある．どの計算方法を選んでも，回帰的クラスタリングの基本処理である，N 個の特徴ベクトルを複数のクラスタに分割するアルゴリズムの実装に大きな影響を与えない．かつ，計算方法の変換によるアルゴリズムの修正は簡単にできる．たとえば，類似性の計算方法がユークリッド距離から相関係数に変わる場合，アルゴリズム中の "最大" と "最小" の判断を逆にすればよい．したがって，以下では単にユークリッド距離を類似性の計算手法としたクラスタリングのアルゴリズムについて述べる．

2.1.1 クラスタリング

分類対象となる特徴ベクトルを $\{X_1, X_2, \cdots, X_N\}$ とする．それらを1つの原始クラスタとして考えて，M 個のクラスタ $\{C_1, C_2, \cdots, C_M\}$ に分割するアルゴリズムは次のようにまとめられる．ここで M は cluster-tree の分岐数であり，その値は実際の応用ニーズに応じて自由に決めることができる．

(1) N 個の特徴ベクトル中から類似性が最も低い2つの特徴ベクトル X'_1, X'_2 をまず選び出し，2つのクラスタを設置する．

(2) 原始クラスタ中に残る特徴ベクトルから，後述する最大距離アルゴリズムに従って，もう1つの特徴ベクトル X'_j $(j \in [3, M])$ を選び出し，1つのクラスタを設置する．

(3) 選び出した特徴ベクトルの数が M に至れば次のステップに行く．そうでなければ，ステップ (2) に戻る．

(4) 残る $N - M$ 個の特徴ベクトルを，それぞれと最も類似する X'_i $(i \in [1, M])$ の属するクラスタに入れる．

2.1.2 最大距離アルゴリズム

適切に N 個の特徴ベクトルを M 個のクラスタに分割するため，すでに M' 個のクラスタ $(M' = 2, \cdots, M-1)$ を設置したとし，$M'+1$ 番目のクラスタを設置する際，以下の最大距離アルゴリズムにより候補を選ぶ．

(1) 残る $N - M'$ 個の特徴ベクトルと M' 個のすでに選び出した特徴ベクトルの距離を計算し，M' 個の距離から最小の 1 個を選び，d_i $(i \in [1, N - M'])$ とする．

(2) 以上の $N - M'$ 個の計算結果 (d_i) から最大の1個を選び出し，それに対応する特徴ベクトルを新しいクラスタの候補にする．

2.1.3 クラスタの中心と半径

クラスタの平均的特徴を表すため，クラスタにある特徴ベクトルの各属性の平均値を基に，クラスタの「中心」(C) と呼ぶベクトルを作る．また，クラスタ中の特徴ベクトルの分布状態を表すために，そのクラスタから C との類似性が一番小さい特徴ベクトル (B_r) を選んで，その類似性値をクラスタの「半径」(R_c) とする．そうすれば，クラスタ中のすべての特徴ベクトルは C と R_c により構成した n 次元特徴空間内の1つの「球」中に分布していると考えられる．後述のように，このような記述を用いれば，排除探索手法中の判断基準を簡単に求めることができる．

2.2 インデックスの構成

以上のクラスタリング手法を再帰的に繰り返せば特徴ベクトルの木構造インデックス (cluster-tree) が構成される．具体的なアルゴリズムを次にまとめる．

(1) 分類対象となる特徴ベクトルを 1 つの原始クラスタとする．論理的にそれは cluster-tree の

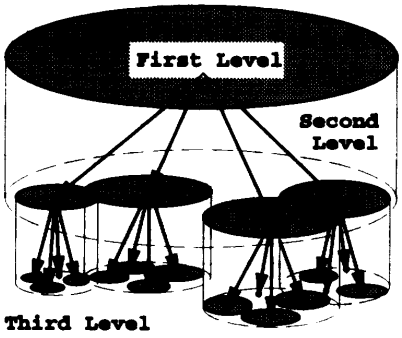

図1 4分岐の cluster-tree の論理構造
Fig. 1 Logical structure of a four-branch cluster-tree.

ルートとなる.

(2) 以上述べたクラスタリングアルゴリズムを用いて，原始クラスタを M 個の小さいクラスタに分割する.

(3) 新たに生成されたクラスタをチェックし，その中に M 個以上特徴ベクトルを持つクラスタをそれぞれ1つの「原始」クラスタとし，ステップ(2)に戻り，もう一度分割する.

分類処理中に生成されたクラスタと cluster-tree の部分木との関係は

(1) 論理的にすべてのクラスタはそれぞれ cluster-tree の1つの部分木と対応する，

(2) あるクラスタを M 個の小さいクラスタに分割すると，M 個の小さいクラスタに対応する部分木はそれぞれ元のクラスタに対応する部分木の1つの分岐にする，

ように記述される．**図1** は4分岐の cluster-tree 論理構造の例を示している.

物理的にはすべての特徴ベクトルを cluster-tree の葉ノードにする．一方，個々の中間ノードは関連のクラスタの記述 C，R_c および B_r から構成され，それぞれ下層の部分木のルートにする．よって，各部分木（1つのクラスタ）中にある特徴ベクトルの平均的特徴および分布状態が cluster-tree の中間ノード中に記述される.

2.3 cluster-tree の特徴

理論的また実際の実験結果によると，回帰的クラスタリングで構成した木構造のインデックスは，ルートから葉へたどって行くと，クラスタ中の特徴ベクトルの数が徐々に減っていくと同時に，クラスタ中の特徴ベクトルの類似性の最小値が高くなるという階層的クラスタリングの分類特徴[10]を持つ．かつ，その分岐数（M）は分類される特徴ベクトルの分布状態に応じたバランスの良い cluster-tree の構築に関して，また二次記憶装置の1ページに格納できる cluster-tree ノードの数によって自由に設定できる.

特徴ベクトル間の類似性を基にデータを分類するクラスタリング手法は具体的な検索サンプルに依存せず，類似性が高いものを1つのクラスタにまとめるという特徴を持つ．しかし，1組の特徴ベクトル間の類似性（仮にその計測関数を Sim とする）関係が数学的類似関係の推移性（transitive）[13]

$$\mathrm{Sim}(x,z) \geq \max_{y}(\mathrm{Sim}(x,y),\mathrm{Sim}(y,z))$$

を満たさないケースが存在するため，相互の類似性が一定値を超える2つの特徴ベクトルを必ずしも同じクラスタにまとめることができない可能性がある．そのような特性を持つクラスタリング手法で構成された特徴ベクトルの木構造インデックス中には，実際の検索に関連する特徴ベクトルと関連しないものが混入し，複数の部分木中に分散保存される現象が存在する．また，cluster-tree の構造は分類対象の特徴空間内の分布状態に依存する．次章では，このような特徴に応じた新たな探索手法について述べる.

3. cluster-tree の探索

前述した cluster-tree の構造的な特徴により，B-tree のような従来の木構造インデックスに用いられる探索手法は cluster-tree に対して適用できない．Croft[4]は特徴ベクトル間の類似性に基づき構成された木構造インデックスの探索問題に対して，検索サンプル Q に関する部分木（クラスタ）C_i のエントロピー $H = -\sum P(C_i|Q)\log P(C_i|Q)$ を用いた探索手法を提案した．しかし，その案には確率 $P(C_i)$ の計算についての理論的な根拠が弱いことや，エントロピーのいき値を決定する一般的な基準を明確に取り上げていないなどの問題点がある.

本提案はこのような問題の解決に応じた新たな探索手法，すなわち，「排除探索」を取り上げる．この手法を用いた探索処理は，cluster-tree の全体を初期「探索範囲」とし，ルートから再帰的な手法で一段下の各部分木をそれぞれチェックし，検索条件と合わないものを探索範囲から除外するように行われる．最後に検索対象は絞られる.

3.1 基本的な方針

すでに述べたように，ある検索サンプルに対して，検索対象となる特徴ベクトルと関係ないものが混入して複数の部分木に保存される可能性がある．しかし，検索対象が cluster-tree 中でどのように分布するかはそれらを見つける前に予想できない．そのため，以下

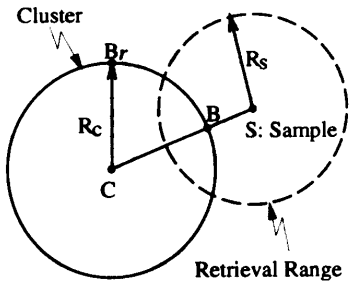

図2 2つの球の特徴空間内の位置関係による判断法
Fig. 2 Judgement criteria based on the position relationship of two spheres.

にこのような問題に対応するcluster-treeの基本探索手法を取り上げる．

特徴ベクトルを用いた類似検索には，検索条件が検索サンプルおよびそのサンプルとの類似度合いの要求という2つの要素からなる．この2つの要素をそれぞれ「中心」と「半径」とすれば，検索条件はクラスタの記述と同様に実際の「検索範囲」と見られるn次元の特徴空間内の1つの「球」を構成すると考えられる．検索サンプルとの類似性値がその半径を超える特徴ベクトルはその検索範囲から外れ，検索条件と合わないことが分かる．したがって，あるクラスタ（cluster-treeの部分木）が検索条件と合うかどうかの判断をその2つの球が交差するかどうかの判断に変換すれば（図2を参照），後述のように排除判断基準は簡単に求められる．すなわち，検索範囲と交差しないクラスタにあるすべての特徴ベクトルは検索範囲から除外できる．

3.2 探索中の判断基準

図2に示しているように，仮に検索サンプルSから判断されるクラスタの中心Cまで直線を1本引くと，それがクラスタ境界とB点のところで交差する．交点Bは実際の特徴ベクトルではなく，判断式を求めるときの中間変数として利用される．クラスタ中心，検索サンプルおよびB点がそれぞれ$C=(c_1,c_2,\cdots,c_n)$，$S=(s_1,s_2,\cdots,s_n)$，$B=(b_1,b_2,\cdots,b_n)$とすれば，それらの間に次の関係がある．

$$b_i=c_i+\lambda(s_i-c_i),\quad (i=1,\cdots,n). \tag{1}$$

ここのλは比率係数（仲介パラメータ）である．求められたλの値から，Sがクラスタの中に入っているかどうかが分かる．

$$\lambda=\begin{cases}\geq 1 & S \text{ in the cluster}\\ <1 & S \text{ out of the cluster}\end{cases}$$

特徴ベクトルの類似性計算に関して，様々な方法[7]が開発されているが，通常，どんな方法を選ぶかは具体的な応用対象によって決められる．異なる類似性の計算方法に対して検索範囲とクラスタ両球の位置関係を判断する具体的な処理方式が違う．ここで様々な研究中で使われているユークリッド距離[2]と相関係数[8]という2つの類似性の計算方法を例として，探索処理中に使用する判断式を求める一般的な手順を取り上げる．

3.2.1 λの計算式

ユークリッド距離： 式(1)とユークリッド距離の定義によって，クラスタ中心Cと仮想的交点Bとのユークリッド距離D_{CB}を式(2)のように求める．

$$D_{CB}=\lambda\sqrt{\sum(s_i-c_i)^2}=\lambda D_{SC}. \tag{2}$$

クラスタはn次元特徴空間内の「球」という記述に従って，D_{CB}とクラスタの半径D_{CB_r}が同等であると考えられる．

$$\lambda=\frac{D_{CB}}{D_{SC}}=\frac{D_{CB_r}}{D_{SC}}. \tag{3}$$

ここのD_{SC}はSからCまでのユークリッド距離である．

相関係数： 2つのベクトル$X=(x_1,x_2,\cdots,x_n)$と$Y=(y_1,y_2,\cdots,y_n)$との相関係数は次の式で計算される．

$$R(X,Y)=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$$

簡潔な表現式を求めるため，ベクトルC, B, Sを$C'=(c'_1,c'_2,\cdots,c'_n)$，$B'=(b'_1,b'_2,\cdots,b'_n)$および$S'=(s'_1,s'_2,\cdots,s'_n)$のような形式に変換する．

$$c'_i=\frac{c_i-\bar{c}}{\sqrt{\sum(c_i-\bar{c})^2}},\quad b'_i=\frac{b_i-\bar{b}}{\sqrt{\sum(b_i-\bar{b})^2}},$$
$$s'_i=\frac{s_i-\bar{s}}{\sqrt{\sum(s_i-\bar{s})^2}}.$$

このような変換によって，元のSとCとの相関係数R_{SC}，CとBとの相関係数R_{CB}およびSとBとの相関係数R_{SB}を新たに導入されたベクトルの内積演算で計算することができる．

$$R_{SC}=\sum c'_i s'_i,\quad R_{CB}=\sum c'_i b'_i,$$
$$R_{SB}=\sum s'_i b'_i.$$

また，以上の変換により，次の2つの式が成り立つ．

$$\sum c'^2_i=1;\quad \sum s'^2_i=1$$

式(1)をCとBとの相関係数の計算式に代入すると，R_{CB}の計算式は次のようになる．

$$\begin{aligned}R_{CB}&=\sum[c'_i+\lambda(s'_i-c'_i)]c'_i\\&=\sum c'^2_i+\lambda\left(\sum s'_i c'_i-\sum c'^2_i\right)\\&=1+\lambda(R_{SC}-1)=R_{CB_r}.\end{aligned} \tag{4}$$

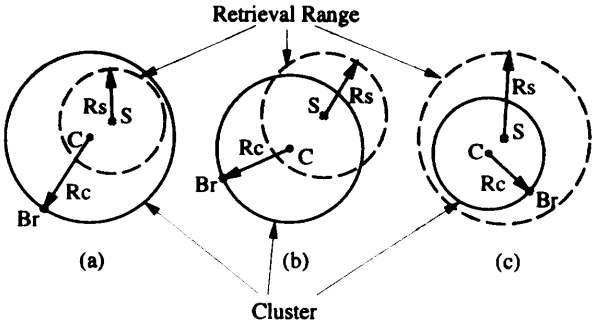

図3 $\lambda > 1$ のときのクラスタと検索範囲との3種類の位置関係
Fig. 3 Types of the position relationships of a cluster and retrieval range when $\lambda > 1$.

クラスタは n 次元特徴空間中の半径= R_{CB_r} である「球」という記述に従い，λの計算式は次のようになる．

$$\lambda = \frac{R_{CB} - 1}{R_{SC} - 1} = \frac{R_{CB_r} - 1}{R_{SC} - 1} \tag{5}$$

ここの R_{CB_r} は相関係数を類似性の計算方法とする場合のクラスタの半径である．

検索サンプルがクラスタの中に入っている場合，検索範囲とクラスタの両球の位置関係はつねに**図3** 中に書いている3種類中の1つである．明らかにその場合クラスタ中の一部の特徴ベクトルが検索対象となる可能性が非常に高い．

3.2.2 部分的交差の判断

たとえ検索サンプルがクラスタの外にあるとしても，そのクラスタ中に検索対象が1個もないとはまだ断定できない．すなわち，クラスタが検索範囲と部分的に交差すれば，交差区域中に特徴ベクトルが存在する可能性がある．そのため，単にλの計算結果を用いてクラスタと検索範囲との位置関係を判断できない場合，他の判断手法を用いることが必要となる．多くの類似性の計算手法に対して，それに応じた判断式を求めるとき，λの計算結果を利用する必要があるため，クラスタが指定された検索範囲と交差するかどうかの判断式は以下に求められる．

ユークリッド距離： 検索サンプル Sと仮想的交点ベクトル Bとのユークリッド距離は $(1-\lambda)D_{SC}$ である．クラスタが検索範囲と交差しない場合

$$(1-\lambda)D_{SC} > R_s \tag{6}$$

という不等式が成立する．式中の R_sはユークリッド距離を類似性の計算法とするときの検索範囲の半径（検索条件中で設定された検索サンプルとの最小ユークリッド距離）である．

相関係数： λの計算結果を用いて，検索サンプル Sと仮想的交点ベクトル Bとの相関係数 R_{SB}を求めると，

$$R_{SB} = \sum s_i'[c_i' + \lambda(s_i' - c_i')] \tag{7}$$
$$= \sum s_i'c_i' + \lambda\left(\sum {s_i'}^2 - \sum c_i's_i'\right)$$

である．さらに変換すれば，

$$R_{SB} = R_{SC} + \lambda(1 - R_{SC}) \tag{8}$$

となる．クラスタが検索範囲と交差しない場合，R_{SB}が指定された検索範囲を超える．すなわち，

$$R_{SC} + \lambda(1 - R_{SC}) < R_s \tag{9}$$

という不等式が成立する．式中の R_sは相関係数を類似性の計算法とするときの検索範囲の半径（検索条件中で設定された検索サンプルとの最大相関係数値）である．

3.3 探索処理の流れ

排除探索手法により，検索対象と無関係な部分木を除外する処理にともない，探索範囲は次第に絞り込まれる．前述したように，ある判断対象のクラスタ（部分木）が検索範囲と交差しない場合，その部分木の中に検索対象はないことが分かる．一方，両方が交差する場合，交差度合いによってその部分木中に，

(1) 検索対象のみ存在する
(2) 検索対象と関係ないものが同時に存在する
(3) 検索対象が存在しない

という3種類の可能なケースがある．したがって，排除探索の処理過程の流れは次のようになる．

(1) クラスタが検索範囲と交差しなければ，その部分木の探索を中止する（排除処理）．
(2) そうでなければ，さらに1段下の各部分木を再帰的に探索し，3つの可能なケースから実際に属するケースが判明する．

4. 性能評価と討論

4.1 インデックス構成

cluster-tree の構成の必要なコストは再帰的クラスタリング手法の特徴によって，次の2つの部分からなる．

M個のクラスタの分割（C_M）

(1) M個のクラスタを設置するコスト：

$$\frac{N(N-1)}{2} + \sum_{M'=2}^{M}(1+M')(N-M')$$

(2) 残る $N-M$個の特徴ベクトルを M個のクラスタにまとめるコスト：

$$(N-M)M$$

2つの部分を合わせて C_Mとなる．

再帰的処理の数（n_c） 分割処理を再帰的に繰り返す数は特徴ベクトルの分布によって異なる．すべてのクラスタの分割結果は同等数の特徴ベクトルを持つM個のクラスタを生成する平均的分布に対して，

$$n_c = \frac{\log N}{\log M} - 1$$

となる．一方，すべてのクラスタの分割結果は$M-1$個の新クラスタが1つの特徴ベクトルしか持たない，残る特徴ベクトルが1つのクラスタにまとまるというcluster-treeのバランスが完全に崩れる分布に対して，

$$n_c = \frac{N-M}{M-1}$$

となる．よって，cluster-treeの構成の総コスト（$C_M \times n_c$）は$O(N^3)$の以下である．

N個の特徴ベクトルを複数のクラスタに分類する処理に対して，K-平均アルゴリズムを用いることもできる．しかし，本提案の回帰的クラスタリング中で採用しているクラスタリング手法と比べ，K-平均の方が最低2倍以上のコストを要する．かつ，一般には分類対象の数が増えれば増えるほど，K-平均の分類コストがさらにアップする傾向がある．K-平均の手法で作った特徴ベクトルの木構造インデックスは，検索コストにおいては回帰的クラスタリング手法と比べると多少は改善されるが，分類コストとを統合して判定すると不利であることが実験で確認された．

4.2 実装テスト

本提案の検索コストのオーダは次の2つの理由により，簡単に表すことができない．

(1) 特徴ベクトルの分布の影響により，同じ数の特徴ベクトルから作ったcluster-treeの形が違うため，検索サンプルごとに検索コストが異なってしまう可能性がある．

(2) 検索範囲の変換により検索対象の数が変わるので，必要な検索コストが異なる．

そのため，本文は線形検索法（LAR）[9]との比較，すなわち，同じテストデータに対して，特徴ベクトルのcluster-treeを構成したうえで排除探索を用いて類似するものをサーチする場合，LARよりその検索コスト（特徴ベクトルをアクセスする数）をどの程度削減できるかという比較を評価の基本手段として採用している．

4.2.1 テスト用のデータについて

232件の英文の全文（経済に関連する記事とAIDS予防に関連する記事の概要）を実験データとして使用した．テキストから特徴ベクトルを抽出する作業には“The American Heritage Dictionary of the English Language”をもとに作った実験用の辞書を使用した．この実験用の辞書は48個の分野に関連する5653個の単語および決まり文句（phrase）があり，各単語と決まり文句はそれぞれ1つ以上の分野情報コードを持つ．英文の固有の特徴を考慮し，辞書から分野情報を検索する際，辞書中に同じ単語で始まる決まり文句が同時にあれば，決まり文句を優先する．以下に具体的な抽出処理手順を取り上げる．

テキスト中の単語または決まり文句が属する分野フィールド（science field）を実験用の辞書から求め，そのテキスト中に使用されている分野フィールドの回数を求める．この回数を分野フィールドの重みという．このようにして求めた重みの大きいフィールドをこのテキストの主要なフィールドであるものとする．テキスト中のある単語あるいは決まり文句が複数の分野フィールドに属する場合，より主要なフィールドをその単語あるいは決まり文句の分野フィールドと定める．このように求められた各分野フィールドの重みを計算し1つのベクトルにまとめ，正規化した後にこのテキストの特徴を表す特徴ベクトルが構成される．

4.2.2 検索コスト

排除探索法を用いてcluster-treeから類似するものをサーチする場合，必要なコストは検索サンプルによって違うケースがあるため，評価実験☆では平均検索コスト（異なる特徴ベクトルを検索サンプルとし類似検索を行ったときのコストの平均値）を用いてLAR手法と比較する．図4は相関係数を類似性の計算手法とした場合，実験中とった両方の検索コストの比較結果を示している．通常は，LAR手法の検索コストは検索範囲と関係なく固定値（$O(N)$）である．比較しやすくするためLAR手法の検索コストを1と見なし，本提案の検索コストをそれとの相対数値で表す．図の中で，RO/TOは検索範囲が変わる（類似度合いの要求が0.5から0.95に変化し，関連する検索範囲が次第に小さくなる）とき，平均呼出し率（異なる検索サンプルで類似検索を行うとき，呼び出されたオブジェクトの数がデータベースにある特徴ベクトルの全体に占める割合の平均値）の変化を表す．図4の中に示しているように，検索範囲が大きくなると，本提案の検索コストは徐々に増える．しかし，呼出し率が10%以下である場合は，本提案の検索コストはLAR手法の

☆ 実験用のプログラムはCで作ったもので，評価テストはSUN IPC/SunOS 4.1.3上で行われた．

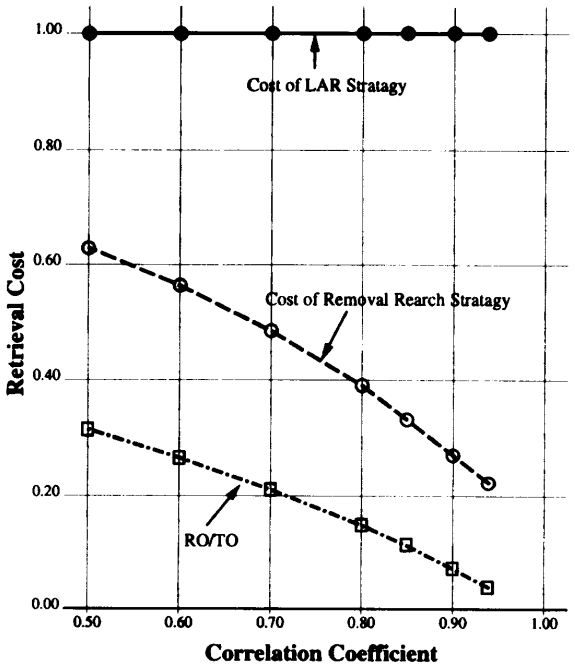

図4 相関係数を類似性の計算手法とした検索コスト
Fig. 4 Retrieval cost when correlation coefficient used as the measuring method of similarities.

検索コストの約3分の1にとどまる☆.

4.3 探索処理効率の改良

ユークリッド距離を類似性の計算方法とする場合，実際の実験結果の分析によって，クラスタと検索範囲との交差度合いの判断基準を次のように調整すれば，探索処理の効率をあげることができる.

4.3.1 実際の経験に基づく改良方法

図5の例を用いて説明すると，クラスタ中に3つの特徴ベクトル B_r, X, Y があり，それらの幾何的位置関係は図中に示すようになっているとする．このような場合，明らかに B_rB_m 線分の右側区域の中に特徴ベクトルが存在しないが，クラスタが検索範囲と交差しているので，式(6)でこのクラスタが排除対象であることはさらに下の部分木にいかないと判断できない．しかし，クラスタの境界を狭くすれば，つまり，CB_r と SC 間の角度 θ が 90° に近づくと，

$$D_{SC} - \left|\frac{(CB, SC)}{|SC|}\right| > R_s$$

を用いて交差度合いを判断することにより，このクラスタが検索条件と合わないことをより早く発見することができる．ただし，θ がどんな数値を超えたら，

☆ 相関係数を類似性の計算手法とした場合の具体的な実験結果をあげると，呼出し率を0.105にするとき，検索1回あたりの平均所要時間は0.216秒であった（このときの検索コストの対LAR比は37%）.

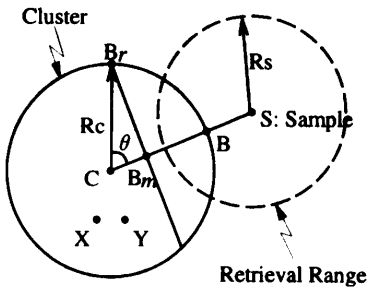

図5 実際の経験に基づくクラスタ境界の修正
Fig. 5 Revision of cluster boundary based on the experiments.

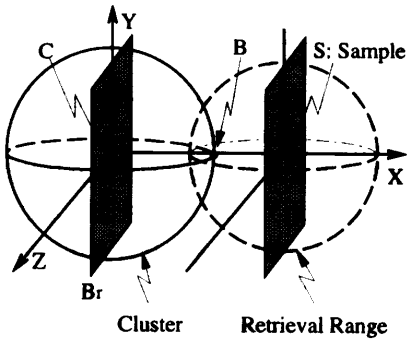

図6 統計手段を用いたクラスタ境界の修正
Fig. 6 Revision of cluster boundary based on statistics.

式(6)の代わりに上式でクラスタと検索範囲との位置関係を判断するかは実際のデータに依存する．式中の $(CB, SC)/|SC|$ はクラスタ中心から B_m までの距離である.

4.3.2 統計手段を用いた改良手法

同じクラスタにあるすべての特徴ベクトルはそれらの一部属性が数字上に同等あるいはかなり近い場合，クラスタ中の狭い領域内に分布する．この現象を図6の三次元の例を用いて示す．図6中に示されている各特徴ベクトルはそれらの横軸方向の属性値が大体同じであるため，ほぼYZ平面に分布している．このような現象が出る場合は，検索範囲とクラスタ両球が交差しても交差区域中に特徴ベクトルは1つもない可能性がある.

このような現象に対して，統計手段を用いてクラスタ境界を再設定すれば，検索条件と合わない部分木をより早く見つけることができる．実験結果によると，この方法によれば検索の効率が2倍に向上できる．新しいクラスタ境界の設定は前のクラスタ中心から一番離れた特徴ベクトル B_r の代わりに，特徴ベクトルの各属性値の標準偏差

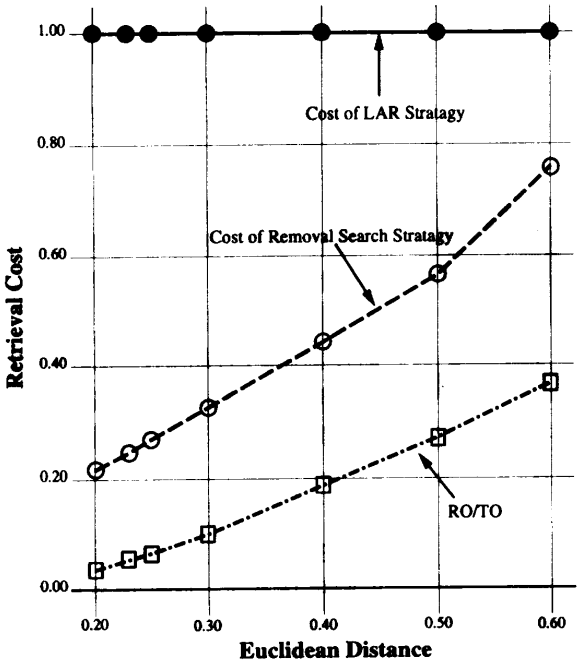

図7 ユークリッド距離を類似性の計算手法とした検索コスト
Fig. 7 Retrieval cost when Euclidean distance used as the measuring method of similarities.

$$b'_{ri} = \sqrt{\frac{1}{FN-1}\sum_{j=1}^{FN}(v_{ij} - \bar{v_i})^2}, \quad i = 1, \cdots, n \qquad (10)$$

で作った新境界ベクトル B'_r を用いる．上式の中に FN は特徴ベクトルの属性の数で，n はクラスタ内の特徴ベクトルの数である．さらに，不等式 (6) の代わりに，

$$D_{SB} = \sqrt{\sum_{j=1}^{FN}(|s_j - c_j| - b'_{rj})^2} > R_s \qquad (11)$$

を用いて，クラスタが検索範囲と交差するかどうかを判断する．図7 はこの改良手法による実験の結果を示している．ユークリッド距離を類似性の計算方法とする場合でも，検索範囲が大きくなるほど本提案のインデックス方法の検索コストは増えるが，呼出し率が10%以下である場合，この方法の検索コストがLAR手法の検索コストの約3分の1以下になる☆．

5. むすび

本論文では，特徴ベクトルを用いたデータベース類似検索において，特徴ベクトルの新しい分類方法である回帰的クラスタリングと関連の排除探索手法を提案した．本提案は対象生データの種類を問わず，特徴ベクトルモデルを用いた類似検索に対して適用できる．いくつかの評価実験結果により本手法の検索コストは，呼出し率が10%以下の場合，通常よく使われる線形検索LAR手法の3分の1程度であることが明らかになった．

本提案の方法をより幅広く利用できるようにするため，様々な類似性の計算手法に対し排除探索手法における判断式を開発していくことが今後の課題と考えている．

☆ ユークリッド距離を類似性の計算手法とした場合の具体的な実験結果をあげると，呼出し率を0.101にするとき，検索1回あたりの平均所要時間は0.207秒であった（このときの検索コストの対LAR比は34%）．

参考文献

1) Bordogna, G. and Pasi, G.: A Fuzzy Linguistic Approach Generalizing Boolean Information Retrieval: A Model and Its Evaluation, *J. Am. Inf. Sci.* pp.70–82 (1993).
2) Chang, T. and Kuo, C.C.J.: Texture Analysis and Classification with Tree-Structured Wavelet Transform, *IEEE Trans. Image Processing*, Vol.2, No.4, pp.429–441 (1993).
3) Gong, Y., Zhang, H., Chuan, H.C. and Sakauchi, M.: An Image Database System with Content Capturing and Fast Image Indexing Abilities, *1994 International Conference on Multimedia Computing and Systems*, pp.121–130 (1994).
4) Croft, W.B.: A Model of Cluster Searching based on Classification, *Inf. Syst.*, Vol.5, pp.189–195 (1980).
5) Gowda, K.C. and Krishna, G.: Agglomerative Clustering Using the Concept of Mutual Nearest Neighborhood, *Pattern Recogn.*, Vol.10, pp.105–112 (1978).
6) Jolion, J.M., Meer, P. and Bataouche, S.: Robust Clustering with Applications in Computer Vision, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.PAMI-13, No.8, pp.791–802 (1991).
7) Kirriemuir, J.W. and Willett, P.: Identification of Duplicate and Near-Duplicate Full-Text Records in Database Search-Outputs Using Hierarchic Cluster Analysis. *Program*, Vol.29, No.3, pp.241–256 (1995).
8) Liddy, E.D., Paik, W. and Yu, E.S.: Text Categorization for Multiple Users Based on Semantic Features from a Machine-Readable Dictionary, *ACM Trans. on Information Systems*, Vol.12, No.3, pp.278–295 (1994).
9) Willet, P.: Document Clustering Using and Inverted File Approach, *Journal of Information Science*, pp.223–231 (1980).

10) Willet, P.: A Note on the Use of Nearest Neighbors for Implementing Single Linkage Document Classifications, *J. Am. Soc. Inf. Sci.* pp.149–152, May 1984.
11) Rijsbergen, C.J.V. and Croft, W.B.: Document Clustering: An Evaluation of Some Experiments with the Cranfield 1400 Collection, *Inf. Process. Manage.*, Vol.11, pp.171–182 (1975).
12) Voorhees, E.M.: Implementing Agglomerative Hierarchic Clustering Algorithms for Use in Document Retrieval, *Inf. Process. Manage.*, Vol.22, No.6, pp.465–476 (1986).
13) Zadeh, L.A.: Similarity Relations and Fuzzy Orderings, *Fuzzy Sets and Applications: Selected Papers by L.A. Zadeh*, pp.81–104. A Wiley-Interscience Publication, John Wiley & Sons (1987).
14) 柴田義孝，藤本道哲：感性検索を可能とするヒューマンインタフェースの性能及び機能評価，第49回情報処理学会全国講演論文集 (4)，pp.165–166 (1994).
15) 大保信夫，石川佳治：シグネチャファイルによる集合値検索のコスト評価，情報処理学会論文誌，Vol.36, No.2, pp.383–395 (1995).
16) 伊藤彰義，遠藤 武，堀桂太郎，島村 徹：階層的印刷漢字認識システムにおける字種を複数クラスタに登録する辞書構成法，電子情報通信学会論文誌，Vol.J78-D-II, No.6, pp.896–904 (1995).
17) 南，中村：画像工学，コロナ社 (1989).
18) 小坂哲夫，松永昭一，嵯峨茂樹：木構造話者クラスタリングを用いた話者適応，電子情報通信学会論文誌，Vol.J78-D-II, No.1, pp.1–9 (1995).

(平成7年10月 3 日受付)
(平成8年 9 月12日採録)

姚 左軍

昭和58年上海交通大学分校・電子卒業．平成4年東京大学大学院電子工学専攻修士課程修了．平成8年同大学院博士後期課程単位取得中退．同年セコム情報システム（株）入社．現在ネットワーク応用開発に従事．電子情報通信学会会員．

濱田 喬（正会員）

昭和16年生．昭和39年東京大学工学部電気工学科卒業．昭和44年同大学院工学系研究科電子工学専攻博士課程修了，工学博士．同年東京大学生産技術研究所助教授．昭和48～49年カリフォルニア工科大学客員助教授．昭和61年学術情報センター教授．昭和62年東京大学工学部教授（併任）．計算機言語，分散処理，オブジェクト指向データベース等の研究に従事．著書「道路交通管制」，「Road Traffic Control」．電気学会，情報処理学会会員．