クラスタリングを用いた楕円体問合せ推定法の改良

1P－9

呉　越勝[†]　石川　佳治[††]　北川　博之[††]

[†]筑波大学 工学研究科 [††]筑波大学 電子・情報工学系

1　はじめに

マルチメディアデータベースの検索を支援するために、ユーザから与えられたサンプルデータより、それらの特徴空間上の分布をもとに、自動的に各特徴量の次元に重み付けし、適切な問合せを推定するようなアプローチが[1]で提案された。この楕円体問合せ推定法は、与えられたサンプルから重みを定めるために楕円体距離関数を推定するアプローチに基づく。この手法では、ユーザから指定された複数のサンプルをもとに、統計的な処理を用いて、ユーザの意図を反映した適切な問合せを自動的に推定できる。しかし、楕円体問合せ推定法では、ユーザから指定されたサンプルデータの集合が一つのクラスタを構成することを仮定していた。この仮定が成立しない場合には、推定結果として得られる問合せがユーザの意図を反映しないものとなってしまうおそれがある。

よって本稿では、ユーザが指定したサンプル集合にクラスタリングを適用することにより、クラスタを検出して問合せの推定に用いる手法を提案する。検出された各クラスタに個別に楕円体問合せ推定法を適用することで、複数のクラスタが存在する場合の問題の改善をはかる。

2　楕円体問合せ推定法

2.1　楕円体距離

本研究で用いる楕円体距離関数は、重み付けしたユークリッド距離を一般化したものであり、

$$\mathcal{D}^2(\boldsymbol{x},\boldsymbol{q})=(\boldsymbol{x}-\boldsymbol{q})^T\mathbf{A}(\boldsymbol{x}-\boldsymbol{q}) \quad (1)$$

のように示される。$\boldsymbol{x}$ と $\boldsymbol{q}$ は d 次元の問合せベクトルと、データベース中のオブジェクトに対応する d 次元ベクトルを示す。ここで'T'は行列の転置を表す。また、行列 $\mathbf{A}$ は正値対称行列であると仮定する。

2.2　楕円体問合せ推定法

以下では、楕円体問合せ推定法[1]の概要について述べる。

ユーザにより N 個のサンプルが与えられたとする。i 番目のサンプルの特徴ベクトルを $\boldsymbol{x}_i = [x_{i1},...,x_{id}]^T$ $(i=1,...,N)$ で表す。また、ユーザは各サンプルに好ましさを表すスコア値 v_i $(v_i>0)$ を指定できるとする。

N 個の例データとスコア値から、ユーザの意図に沿った最良の距離行列 $\mathbf{A}$ と問合せ点 $\boldsymbol{q}$ を推定するため、以下の式の最小化を行う。

$$\min_{\mathbf{A},\boldsymbol{q}}\sum_{i=1}^{N}\mathcal{D}^2(\boldsymbol{x}_i,\boldsymbol{q})=\min_{\mathbf{A},\boldsymbol{q}}\sum_{i=1}^{N}v_i^2(\boldsymbol{x}_i-\boldsymbol{q})^T\mathbf{A}(\boldsymbol{x}_i-\boldsymbol{q}) \quad (2)$$

$\mathbf{A}$ が正則という仮定のもとで、上式を最小にするような行列 $\mathbf{A}_{opt}$ と問合せベクトル $\boldsymbol{q}_{opt}$ が次のように得られる。

$$\boldsymbol{q}_{opt}=\overline{\boldsymbol{x}}=\frac{\sum_{i=1}^{N}v_i^2\boldsymbol{x}_i}{\sum_{i=1}^{N}v_i^2} \quad (3)$$

$$\mathbf{A}_{opt}=det(\boldsymbol{\Sigma})^{1/d}\boldsymbol{\Sigma}^{-1} \quad (4)$$

$\boldsymbol{\Sigma}$ は式（5）のような重みづけした共分散行列である。

$$\boldsymbol{\Sigma}=\sum_{i=1}^{N}v_i^2(\boldsymbol{x}_i-\overline{\boldsymbol{x}})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T \quad (5)$$

3　楕円体問合せ推定法の問題点

楕円体問合せ推定法は、ユーザの好みのデータが、特徴空間で一つのクラスタを構成することを仮定していた。この推定法で推定された問合せが次第にこのクラスタに空間的に収束して、検索が成功となる。しかし、図1のようにユーザの要求に適合するデータが二つのクラスタを構成する場合、楕円体問合せ推定法では、図1(a)のように単一の楕円体問合せが生成され、その形状は一般に細長い楕円となる。このような問合せが推定されると、次の問合せ処理のステップでは、問合せの中心 $\boldsymbol{q}$ から、推定された楕円体距離に基づく近傍探索が行われ、$\boldsymbol{q}$ に近い順に k 個のデータが検索されユーザに提示される。たとえば、図1(b)では、'X'で示したデータがユーザに提示されることになる。しかし、図のようなクラスタ構造がある場合、これらの'X'で示されたデータがユーザの要求に合ったデータでない可能性が大きい。楕

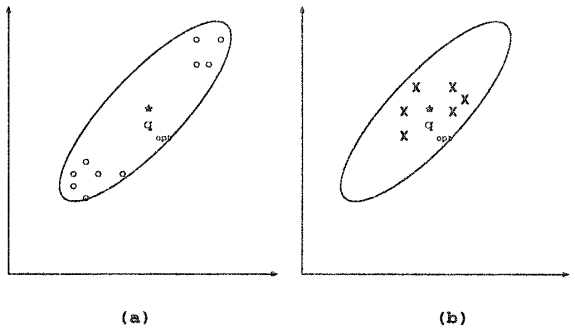

図 1: 楕円体問合せ推定法の問題

円体問合せ推定法は、このような複数のクラスタが存在する場合を考慮していないため、図1で見られるように、ある程度以上は問合せを改善することができないという欠点を有している。

4　楕円体問合せ推定法の改良

本研究では、上記の問題を解決するために、クラスタリングを導入して、ユーザが意図する適切な数のクラスタを抽出することで、楕円体問合せ推定法を改善する。図2に、3節の例における、二つのクラスタを抽出したの状況を示す。クラスタリングを導入した楕円体問合推定法を説明する前に、まずクラスタリング手法について述べておく。現時点で使っているアルゴリズムはK-means法の一種であり、クラスタの数は M であり、以下に示すようにユーザとの対話的処理により適切な数に設定する。詳細なステップは下記のようになる。

1. ユーザが指定したサンプル集合から、互いに十分離れた M 個のサンプルを選ぶ。

Use of Clustering for the Elliptical Distance Estimation Method

Yuesheng Wu[†], Yoshiharu Ishikawa[††], Hiroyuki Kitagawa[††]

[†]Doctoral Degree Program in Eng., Univ. of Tsukuba

[††]Institute of Info. Sci. and Elec., Univ. of Tsukuba

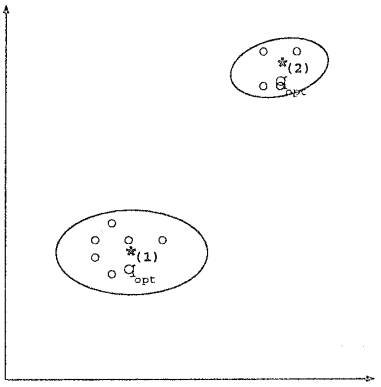

図 2: 楕円体問合せ推定法の問題

(a) ランダムにM個のサンプルを選び、互いの間の距離を調べる。

(b) これを数回繰り返し、最良の M 個のサンプルを選ぶ。

2. 1 で選んだ M 個のサンプルを初期のクラスタ代表として reallocation 法 [2]。を用いる

Reallocation 法

1. M 個の初期クラスタ代表（クラスタの重心となる）を選ぶ。
2. サンプルデータのそれぞれについて
   (a) M 個のクラスタの中から、最もサンプルデータに近い重心をもつものを見つける。
   (b) サンプルデータを選ばれたクラスタに割り当てる。
3. M 個のクラスタのそれぞれについて、重心の再計算を行う。
4. 各クラスタの内容が変化しなくなるまで 2, 3 のステップを繰り返す。

4.1 問合せ処理ステップ

楕円体問合せ推定法の処理を以下のように改良する：

初期フェーズ（改良前と同じ）

1. サンプル集合 S をユーザに提示し、ユーザはサンプル集合の中から好みのサンプルを複数個選択
2. 選ばれたサンプルをもとに、楕円体問合せ推定法により問合せ q_{opt} と $\mathbb{A}_{opt}$ を推定
3. データベースに問合せ Q を発行し、類似度の順に上位 K 個のデータを検索
4. K 個のデータを次のサンプル集合とし、繰返しステップへ

繰返しフェーズ (改良版)

1. クラスタ数 M := 1 とする。
2. サンプル集合 S (M 個のクラスタからなる) をユーザに提示。
3. ユーザが好む新たなサンプルが与えられたサンプル集合中に存在したとき:
   (a) 適合率（適合サンプル数／K）$> \alpha$ ならば、停止（α は定数）。
   (b) 新たに選ばれたサンプルを加えたサンプル集合を、M 個のクラスタに再クラスタリングする。
4. ユーザが好む新たなサンプルが存在しなかったとき:
   (a) M := M + 1 とする。
   (b) M> β （β は定数）となったら、停止。
   (c) 現在のサンプル集合を M 個のクラスタに再クラスタリングする。
5. 各クラスタ i について、楕円体問合せ推定法により問合せ位置と $q_{opt}^{(i)}$ 距離行列 $\mathbf{A}_{opt}^{(i)}$ を推定。
6. データベースに各問合せ Q_i を発行し、類似度の順に上位 $K(= K_1 + ... + K_M)$ 個のデータを検索。
7. K 個のデータを次のサンプル集合 S とし、ステップ 2 に戻る。

繰返しフェーズにおいてユーザが好ましいサンプルを新たに発見できなかった場合、ステップ 4 によって、クラスタの数 M を 1 増やして再クラスタリングを行なう。クラスタの数があまりに多くなる（定数 β より大きい）と、推定が失敗と考え、処理を終了する。

5 実験概要

本研究では、楕円体問合せ推定法の改善案を提案した。実験において、本改善案と楕円体問合せ推定法を比較し、本改善案の有効性を評価することは、実験の一つの目的である。

実験のために、楕円体問合せ推定法と本改善案を用いる類似画像検索システムを構築する予定である。実験には、色ヒストグラムとテクスチャ（粗さ、コントラスト、方向性）[3] を画像の特徴として利用する。実験用画像は、テクスチャ特徴を有する実画像（岩石、芝、木の皮など）を利用している。現在、複数の特徴の抽出と並行して、実験作業を進めている。

6 まとめ

本研究では、楕円体問合せ推定法の改善として、ユーザが与えられたサンプルの特徴空間上で単一のクラスタ構成する仮定に必ずしも満たされない場合にクラスタリングを用いた改善案を提案した。今後の課題として、実験を行い、本改善案を評価、及び改良することである。

参考文献

[1] Y. Ishikawa, R. Subramanya, and C. Faloutsos: "MindReader: Querying databases through multiple examples", *in Proc. of VLDB*, pp. 218–227, New York, NY, Aug. 1998.

[2] W. B. Frakes and R. Baezo-Yates(eds.): *Informaation Retrieval Data Structures & Algorithms*, Prentice-Hall, Englewood Cliffs. NJ, 1992.

[3] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki: "Textural feathures corresponding to visual perception", *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-8(6), June 1978.