

MindReader: 例にもとづくデータベース問合せ手法

石川佳治[†] Ravishankar Subramanya[‡] Christos Faloutsos*

[†] 奈良先端科学技術大学院大学 情報科学研究科

ishikawa@is.aist-nara.ac.jp

[‡] ピッツバーグスーパーコンピューティングセンター

* カーネギーメロン大学

マルチメディアデータベースの内容にもとづく検索では、通常、特徴量による類似検索処理が行われるが、ユーザが自身の意図に合うように特徴量の組合せの指定をすることは容易なことではない。そこで本稿では、例にもとづく問合せ手法 *MindReader* を提案する。ユーザは複数の例と、例のそれぞれがどの程度ユーザの意図に沿っているかを反映したスコアを提供する。提案手法は、提供されたデータからどの特徴や特徴間の相関が重要であるかを推定し、ユーザの意図を反映した距離関数と理想的な問合せ位置を出力する。この問合せ位置と距離関数による問合せを行うことで、ユーザは自身の意図に合った問合せ結果を得ることができる。本稿ではこの手法の理論の概要と実験結果について報告する。

MindReader: Querying databases through multiple examples

Yoshiharu Ishikawa[†] Ravishankar Subramanya[‡] Christos Faloutsos*

[†] Graduate Institute of Information Science, Nara Institute of Science and Technology

[‡] Pittsburgh Supercomputing Center

* Carnegie Mellon University

In this paper, we provide a user-friendly, but theoretically solid, example-based query processing/refinement method called *MindReader*. We allow the user to give several examples, and optionally, their 'goodness' scores. Based on the examples and scores, our method "guess" which features and correlations are important.

Our contributions are twofold: (a) we formalize the problem as a minimization problem and show how to solve for the optimal solution, completely avoiding the ad-hoc heuristics of the past. (b) we are the *first* that can handle 'diagonal' queries. Experiments on synthetic and real datasets show that our method estimates quickly and accurately the 'hidden' distance function in the user's mind.

1 はじめに

マルチメディアデータベースをはじめとする現代のデータベースにおいては、距離や類似度にもとづく問合せがより重要となってきた。しかし、データベースのユーザにとって、与えられた特徴 (例: 色のヒストグラム, 形状) のもとで問合せを適切に表現することは必ずしも容易ではない。例として、画像データベースにおける夕日の画像の検索を考える。QBIC (Query By Image Content) [FBF+94], Virage [Vir] のような既存の画像検索システムの多くでは、複数の特徴を組み合わせた画像検索が可能となっている。このようなシステムでは、ユーザの問合せ作成を助けるため、Query by (Visual) Example [HK92] の機能や、ユーザの特徴に関する好みを獲得するためのインタフェース (例: スライドバー) が提供されている。これにより、ユーザは夕日の画像やスケッチを例として画像検索システムに提供し、スライドバーを用いて色の特徴に高い重要度を、形状に中間的な特徴を設定し、検索したい画像を指定することができる。しかし、このような問合せ作成方式は単純な問合せにしか適用することができない。より複雑な問合せにおいて、それぞれの特徴にどのように相対的な重要度を与えたいかを決定することは、ユーザにとって容易ではない。

同様の状況は従来のデータベースにおいても現われる。リレーショナルデータベースシステム上に構築された VAGUE システム [Mot88] では、属性のドメイン上にメトリクス (data metrics) の概念を、属性値の比較のために類似比較 (similar-to) 演算子を許すことにより、曖昧な問合せの機能を支援している。類似比較に用いられる類似度の尺度をユーザから与えられた例をもとに動的に導出できるなら、この曖昧問合せの機能を有効に活用できることになる。このアイデアを示すために、医療記録のデータベースへの例にもとづく問合せを考える。「太り気味の人」を検索するために、ユーザがデータベースにいくつかのサンプルを与えたとする。この様子を図 1 に示す。ここでは、ユーザの意図を反映するために、各サンプルに対し重要度が指定できることを想定している。サンプルの散らばり具合から、ユーザが体重/身長のある範囲内にある人を求めていることが見てとれる。このようなユーザの意図を反映した類似度を自動的に導出できれば、例にもとづく類似検索が従来のデータベースにおいても実現できることになる。

図 1 には、この論文において重要な二つの概念が現われている。

- 斜交問合せ: 上の例で示したように、与えられた複数の例から、それらを反映した類似度 (あるいは距離尺度) を導いた場合、その類似度や距離にもとづく問合せの検索範囲は円とはならず、一般には楕円となる。このような一般化した類似度や距離にもとづく問合せ

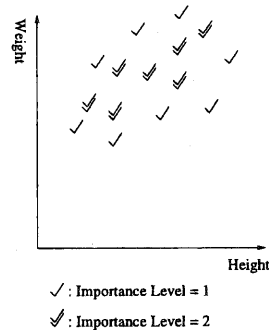


図 1: 例にもとづく問合せ

せを、ここでは斜交問合せ (diagonal query) と呼ぶ。

- 複数レベルのスコア: 与えられた例がどの程度ユーザの意図に合っているかを、複数のレベルで指定できるものとする。類似度や距離尺度の導出では、ユーザが指定したレベル値を反映することが求められる。

上で述べたように、本論文では、ユーザにより与えられた複数の例、およびそれぞれの例に対しユーザが与えたスコアをもとに、ユーザが意図している類似度 (距離尺度) を自動的に導出する手法 MindReader について述べる。検索結果のデータに対しさらにユーザがスコアづけを行い、提案手法を再び適用すれば、類似度 (距離尺度) の推定結果がさらに改善されると考えられる。このような検索とフィードバックの繰返しにより、ユーザは特徴の組合せなどの詳細に踏み込まずに対話的な検索を行うことができる。

2 関連研究

情報検索やデータベースの分野では、ユーザから与えられた例を用いて問合せを行ったり、問合せ結果に関するユーザの判断をフィードバックすることにより問合せを改善しようとするのが試みられてきた。そのようなアプローチは、おおまかに、(a) 問合せ位置の移動、および (b) 再重みづけという二種類に分類できる。これらの概念は直交するものであるが、以前の研究ではこれらを組み合わせたアプローチはあまり見られなかった。

2.1 問合せ位置の移動

この手法は、直観的には、与えられた例のうち問合せに適合 (relevant) していると判断されたものの方向に向かうように、一方、不適合 (irrelevant) と判断された例から離れるように問合せ自体の位置を移動させていくという手法である。このようなアイデアで問合せを洗練するアプローチは、情報検索の分野では適合フィードバック (relevance feedback) と呼ばれている [Har92, SL96]。

たとえば、ベクトル空間モデルにもとづき、有名な適合フィードバック方式の一つである Rocchio の式は以下のように与えられる [Roc71].

$$Q_1 = Q_0 + \beta \sum_{i=1}^{n_1} \frac{R_i}{n_1} - \gamma \sum_{i=1}^{n_2} \frac{S_i}{n_2} \quad (1)$$

ここで Q_0 は最初に与えられた問合せベクトルであり、 R_i は適合文献 i に対する文献ベクトルであり、 S_i は不適合文献 i に対する文献ベクトルである。 n_1, n_2 は、それぞれ適合文献数、不適合文献数である。適合フィードバックの結果として、問合せ位置は Q_0 から Q_1 に移動するととらえることができる。この問合せ位置を移動させるアプローチは、画像検索の分野においても、MARS (Multimedia Analysis and Retrieval System) システムにより試みられている [RHM97, RHM98]。彼らが $tf \times idf$ と呼ぶ手法では、画像の特徴ベクトルから疑似的な文献ベクトルが生成され、Rocchio の式が直接的に適用される。

2.2 再重みづけ

第二のアプローチは再重みづけ (re-weighting) にもとづく手法である。たとえば上記の MARS システムでは、 $tf \times idf$ 法と別個に標準偏差法 (standard deviation method) と呼ばれる問合せの洗練手法が提案されている。このアイデアは非常に明解なもので、各データが n 次元の特徴空間上の点として表されることを前提として

この手法では、まずデータオブジェクトの集合に対し、ユーザによる適合性の判定 (relevance judgment) が行われる。次に、適合と判定されたデータを各次元に対し写像する。 j 番目の次元 ($1 \leq j \leq n$) に写像されたデータの散らばり (標準偏差) が大きければ、 j 番目の次元の値についてはユーザはどの値でも構わず、さほど興味をいだいていないことが想像できる。そこで、その j 番目の次元には小さい重み w_j を与えることにする。逆に散らばりが小さい次元については大きい重みを与えることにする。この考え方にもとづいて、MARS システムでは $w_j = 1/\sigma_j$ ($1 \leq j \leq n$) という重みの設定が行われ、重みつきユークリッド距離 (後述) により距離が計算される。しかし、 $1/\sqrt{\sigma_j}$ や $1/\log(\sigma_j)$ などの他の重みづけの選択肢と比べ、 $1/\sigma_j$ の重みづけが優れているかどうかについては、MARS システムの論文では特に議論はなされていない。

提案する手法 (MindReader) は以上の二種類の問合せの洗練手法を特殊な場合として含んでいる。さらに、

1. アドホックなヒューリスティック (Rocchio の式における β や γ など) を用いない。
2. 複数レベルのスコアを扱える。
3. 斜交問合せを扱える唯一の手法である。

という点が特徴である。

3 提案手法

3.1 基本的なアイデア

本研究で想定しているユーザとシステムの対話のステップは、たとえば以下になる。

1. ユーザは MindReader システムに対し求めたいデータの例 (一般に複数個) を提示する。それぞれの例には、ユーザによるスコア (0, 1 の 2 レベルあるいは複数レベル) を付与する。
2. MindReader システムは、提示された例とスコアから、ユーザが意図している距離関数と問合せの中心位置を推定し、推定結果にもとづいてデータベースに問合せを発行する。
3. データベースシステムから得られた結果はランクづけされ、たとえば上位から 10 件という形でユーザに提示される。ユーザは、結果に満足すれば問合せを終了し、そうでなければ現在提示されている問合せ結果を新しい例の集合として、ステップ 1 に戻る。

1 節では、マルチメディアデータベースやリレーショナルデータベースを例にあげたが、以降では一般的に、 n 次元の空間データベースを想定して議論を行う。よって、以下の議論における「次元」は、マルチメディアデータベースにおける「特徴」、リレーショナルデータベースにおける「属性」などに対応する。また、空間データベースの立場で議論することから、以下では $0 \leq d \leq \infty$ の範囲の値をとる距離によりデータ間の関連をとらえる。しかし、情報検索などのアプリケーションでは、距離ではなく類似度が用いられることが多い。類似度は通常 $0 \leq s \leq 1$ といった範囲をとり、オブジェクトどうしが類似しているほど値が大きくなる。このような類似度のデータを扱う際には、 $s = \exp(-\frac{d^2}{2})$ などという式により距離と類似度の間の相互変換を行う。このことにより、たとえ類似度でデータ間の関係が与えられたとしても、以下の成果を適用することができる。

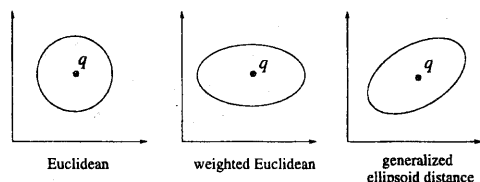


図 2: 距離関数の等距離 (楕) 円

本論文で扱う距離について説明するために図 2 を示す。通常のユークリッド距離の場合、ある点から等距離の点を結ぶと円になる。MARS システムで用いられた、重みづけしたユークリッド距離では、等距離の点は楕円形

状となるが、その楕円の主軸は座標軸に並行なものに限られる。一方、ここで扱う距離は、一般化した楕円距離 (generalized ellipsoid distance) [SK97] と呼ばれ、楕円の主軸は必ずしも座標軸に沿ってはいない。

3.2 手法

表 1 では、以下の議論で用いる記号を示す。ここで用いる距離関数は

$$D(\vec{x}, \vec{q}) = (\vec{x} - \vec{q})^T \mathbf{M} (\vec{x} - \vec{q}), \quad (2)$$

と表される。これは

$$D(\vec{x}, \vec{q}) = \sum_j^n \sum_k^n m_{jk} (x_j - q_j)(x_k - q_k), \quad (3)$$

と等価である。 $\vec{q} = [q_1, \dots, q_n]^T$ は理想的な問合せ点を表す n 次元のベクトルであり、 $\vec{x} = [x_1, \dots, x_n]^T$ はデータエントリに対応するベクトルである。 T は行列の転置を表している。 $n \times n$ 行列 $\mathbf{M} = [m_{jk}]$ は、一般化された楕円距離を規定する対称行列である ($m_{jk} = m_{kj}$)。

表 1: 記号とその定義

記号	定義
n	ベクトルの次元数
N	例の総数
$\vec{x} = [x_1, \dots, x_n]^T$	例ベクトル
$\vec{x}_i = [x_{i1}, \dots, x_{in}]^T$	i 番目の例ベクトル
$\vec{q} = [q_1, \dots, q_n]^T$	理想的な問合せの点
$\mathbf{M} = [m_{jk}]$	距離関数を与える行列
$D()$	距離関数
$\mathbf{X} = [\vec{x}_1, \dots, \vec{x}_N]^T$	N 個の例データを保持する行列
$\vec{v} = [v_1, \dots, v_N]^T$	スコアを保持する行列
$\bar{\vec{x}} = [\bar{x}_1, \dots, \bar{x}_n]^T$	N 個の例データの 重みつき平均の結果
$\mathbf{C} = [c_{jk}]$	N 個の例ベクトルより得られる (重みづけした) 共分散行列
σ_j^2	j 番目の次元に関する 例ベクトルの値の分散

$\vec{x}_i = [x_{i1}, \dots, x_{in}]^T$ を i 番目の例データを表すベクトルとする ($i = 1, \dots, N$)。 $\mathbf{X}$ により $N \times n$ 行列 $\mathbf{X} = [\vec{x}_1, \dots, \vec{x}_N]^T$ を表す。 ユーザがそれぞれの例 $\vec{x}_i$ に対して与えたスコア値を v_i とする。 $\vec{v}$ により、ベクトル $\vec{v} = [v_1, \dots, v_N]^T$ を表す。

本手法の目的は、ユーザが選択した N 個の n 次元の点を用いて、

- ユーザが意図している距離関数の係数、すなわち距離行列 $\mathbf{M}$
- 最適な問合せ点 $\vec{q}$

を推定することにある。推定が正しければ、 $\vec{q}$ を中心として $\mathbf{M}$ で規定される距離で最近隣検索を行うことが、ユーザにとって最良の選択となる。

3.3 定理

与えられた N 個の例データとスコアから、ユーザの意図に沿うという意味で「最良の」距離行列 $\mathbf{M}$ と問合せ点 $\vec{q}$ を見出すためには、どのように最良であるべきかを明確にする必要があるが、ここではペナルティを最小化すると考える。それぞれの例データ $\vec{x}_i$ に対するペナルティが、理想的な問合せ点 $\vec{q}$ からの距離と、ユーザから与えられたスコア v_i の大きさに比例して増えたとし、全体のペナルティはそれぞれの例データに対するペナルティの総和と考えれば、ペナルティの最小化の問題は

$$\min_{\mathbf{M}, \vec{q}} \sum_{i=1}^N v_i (\vec{x}_i - \vec{q})^T \mathbf{M} (\vec{x}_i - \vec{q}) \quad (4)$$

と表現できる。ただし、

$$\det(\mathbf{M}) = 1 \quad (5)$$

という制約を置くことにする。 $\det(\mathbf{M})$ は行列 $\mathbf{M}$ の行列式を表す (このような制約がなければ零行列 $\mathbf{O}$ が最小値を与えることになる)。

式 (5) の制約のもとで式 (4) を最小化できれば、そのときの $\mathbf{M}$ 、 $\vec{q}$ が求める距離行列と問合せ点である。この最小化の問題は、ラグランジュの未定乗数法 (Lagrange multipliers) を用いて解くことができる。以下の定理の詳細は [ISF98] に譲り、ここでは結果のみを示す。

定理 1 データベクトル s の ($\vec{v}$ で重みづけした) 平均を

$$\bar{\vec{x}} = [\bar{x}_1, \dots, \bar{x}_n]^T = \frac{\mathbf{X}^T \vec{v}}{\sum_{i=1}^N v_i}$$

とする。すなわち、

$$\bar{x}_j = \frac{\sum_{i=1}^N v_i x_{ij}}{\sum_{i=1}^N v_i} \quad (j = 1, \dots, n)$$

である。新しい問合せ位置は $\vec{q} = \bar{\vec{x}}$ により与えられる。■

N 個の例ベクトルから得られる、(重みづけした) 共分散行列 (covariance matrix)

$$c_{jk} = \sum_{i=1}^N v_i (x_{ik} - \bar{x}_k)(x_{ij} - \bar{x}_j) \quad (6)$$

を $\mathbf{C} = [c_{jk}]$ とする。次の定理は求める行列 $\mathbf{M}$ を与える。

定理 2 C^{-1} が存在するなら、式 (4) を最小化する M は、

$$M = (\det(C))^{-\frac{1}{2}} C^{-1}. \quad (7)$$

で与えられる。 ■

j 番目の次元に関する、列ベクトルの値の (重みづけした) 分散を

$$\sigma_j^2 = \sum_{i=1}^N v_i (x_{ij} - \bar{x}_j)^2 \quad (8)$$

で定義する。標準偏差法 (MARS の手法) が⁵, MindReader の手法の特殊な場合であることを示す。

定理 3 行列 M を対角行列に制限すると、最良の M は

$$m_{jj} \propto \frac{1}{\sigma_j^2} \quad (9)$$

で与えられる。 ■

すなわち、 M を対角行列に制限するならば、MARS の重みづけは最適なものである。しかし、4 節の実験により、標準偏差法は一般化された楕円距離関数を推定できないことが明らかになる。

ここで一つ重大な問題がある。共分散行列が特異 (singular) であり、逆行列が求められない場合への対処である。本研究の場合、この状況は与えられた例の数が特徴の次元数より小さいとき ($N < n$) に生じる。この問題については、Moore-Penrose inverse matrix (もしくは pseudo-inverse matrix) [GV96] を用いることで解決はかれると考えている [ISF98]。

4 実験

ここでは、人工的なデータと実データに対する実験について述べる。明らかにしたい点は以下のとおりである。

- MindReader は斜交問合せをどのくらい速く、うまく学習できるか。標準偏差を用いる手法と比べどの程度優れているか。
- 問合せ位置の移動について：理想的な問合せ点をすみやかに推定できるか。
- 実データについて、この手法はどの程度有効か。

4.1 準備

評価尺度 ユーザが意図する距離関数の推定における MindReader の性能を評価するため、ここでは二つの評価尺度を導入する。

第一の尺度である CD- k 尺度 (CD- k metric) について述べる。まず、推定した距離関数にもとづいて、問合せ点から近い順に k 個の点を検索する。次に、それら k 個の点のそれぞれについて、今度はユーザが意図している距

離関数により、問合せ点からの実際の距離を計算し、それらの総和をとる。この尺度を CD- k 尺度と呼ぶ。以下では $k = 20$ で実験を行っている。

CD- k 尺度は、アイデアは明解であるが、二つの微妙に異なる距離関数が存在する場合に、それらの違いを尺度の値の差として必ずしも反映できないという欠点をもつ。このような状況は、双方の距離関数を用いて検索した最近隣の k 個の点の集合がまったく同一となる場合に生じる。このような場合においても推定された距離関数の優劣を決定できるようにするため、第二の尺度である MN 尺度 (MN metric) を導入する。MN 尺度は行列のノルムの概念にもとづいており、推定された距離関数を表す行列 M と、ユーザが意図している距離関数を表す行列 M_{hidden} がどの程度異なっているかを測るものであり、

$$\|M - M_{\text{hidden}}\|_2 \quad (10)$$

と表現できる。

スコアの計算 MindReader では、問合せ時にユーザがいくつかの例を提示し、それぞれの例に対し、その良さを示すスコアを付与することを想定している。しかし、後述の実験においては、ユーザが個別に主観的なスコアを付与していくことは、実験の設定の難しさや実験結果の客観性などの面で問題がある。そこで、ここでは距離情報からスコアを自動的に計算する手法を示す。これは以下の実験の一部で使用している。

ユーザが心に描く最良の問合せ点 $\bar{q}$ と、与えられた点 $\bar{x}$ の間の実際の距離は、ユーザが意図する距離関数にもとづいて $d = \mathcal{D}(\bar{x}, \bar{q})$ と与えられる。この距離 d は、先に示した式 $s = \exp(-\frac{d^2}{2})$ により類似度に変換することができる。ここでさらに、 $v = \log \frac{s}{1-s}$ により類似度をスコアに変換する ($-\infty < v \leq \infty$)。このようにして、距離情報からスコアを導くことができる。ユーザが意図する距離において $\bar{x}$ が $\bar{q}$ に近いほどスコアが大きくなることになる。以下の実験においては、推定すべき、ユーザが意図する距離関数は事前に知られているため、この方法でスコアを計算することができる。

4.2 斜交問合せ

第一の実験として、図 3(a) に示す二次元の正規分布データを用いた。このデータの標準偏差は 1 で、平均は (0, 0) である。この実験の目的は以下の二点である。

1. 検索とフィードバックの繰返しによりユーザの距離関数に到達するまでの収束の速さを調べる。
2. MindReader と標準偏差法との性能を比較する。

問合せ点は最初から最適な点 $\bar{q} = (0, 0)$ に置くものとし、この実験では問合せ点の移動については考慮しない。また、この実験ではスコアを 0, 1 の二値に制限している。

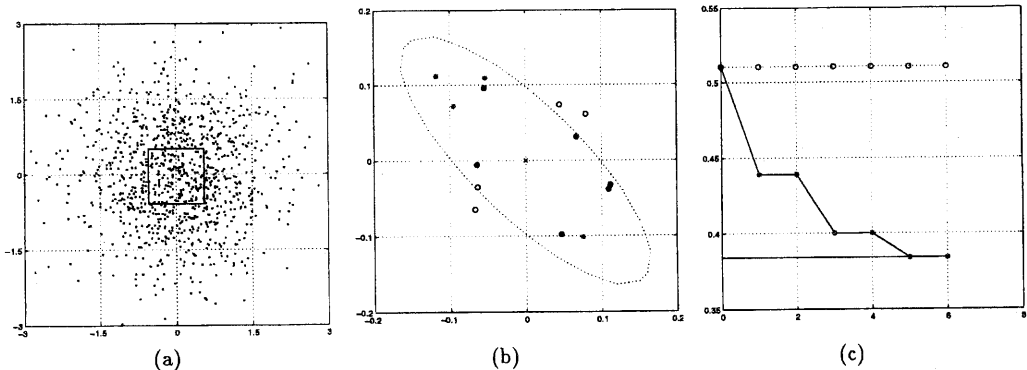


図 3: 斜交問合せの実験: (a) データ集合, (b) ユーザの距離関数にもとづく等距離楕円, および 10 番目までの最近隣点 (*: MindReader, o: 標準偏差法), (c) 収束の速さ: x 軸は繰返し回数, y 軸は $CD-k$ 尺度の値

図 3(b) は, 図 3(a) の正方形で囲んだ部分に対応しており, 検索とフィードバックを繰り返して行い, 収束した時点の状況を示している. 図には, MindReader と標準偏差法によりそれぞれ推定された距離関数にもとづく最近隣の 10 個の点 (それぞれ*, 'o' で表現) を示している. 図中の楕円は, ユーザの距離関数で等距離となる点をつないでできるものである. MindReader で選ばれた最近隣点の集合は図の等距離楕円に沿ったものとなっており, 標準偏差法に比した MindReader の推定の良さが見てとれる.

収束の速さに関する実験結果を図 3(c) に示す. この図では, 各繰返しにおける, MindReader と標準偏差法の $CD-k$ 尺度の値 (すなわち, 最近隣と判断された $k = 20$ 個の点の実際の距離の総和) を示している. 下の水平なグラフは, この実験において達成しうる最良の $CD-k$ 尺度値 (すなわち, 総和が最小になる k 個の点の組合せに対する $CD-k$ 尺度値) を示している. 一方, 'o' で表している上のグラフは標準偏差法を表している. MindReader は*で表しており, この実験では, 5 回の繰返しで最良の尺度値に到達できた. この結果により, 与えられたユーザの距離関数を正しく推定できたことになる. 一方, 標準偏差法は傾斜した距離関数には対応できなかった.

図 4 は, 検索とフィードバックを繰り返すことによる問合せの洗練のようすを示している. 図では, ユーザの距離関数の等距離楕円を点線で, MindReader により推定された等距離楕円を実線で表している. 標準偏差法で推定された距離関数は図中の円に対応している.

問合せが出された直後の初期状態 (繰返し 0 回目) を図 4(a) に示す. 問合せ開始時点では, ただ一つの点 $\bar{q} = (0, 0)$ を例として与えている. 例が一つだけであるため, 0 回目の繰返しでは MindReader, 標準偏差法ともに等距離の点は円の形状をなす. しかし, 検索とフィードバックの繰返しが進むにつれ, MindReader の方はユーザの距

離関数に収束していく. 一方, 標準偏差法では繰返しによる改善はみられない.

4.3 問合せ中心の移動

これまでの実験では, 理想的な点 $(0, 0)$ から問合せを開始していた. 図 5 は, 定理 1 により, MindReader が最良の問合せ中心を推定して問合せ位置を移動できる様子を示している. 実験では, 最初の例を点 $(0.5, 0.5)$ に与えた. 2 回の繰返しで正しい問合せ中心に達し, 6 回の繰返しでユーザの距離関数にほぼ収束することがわかる.

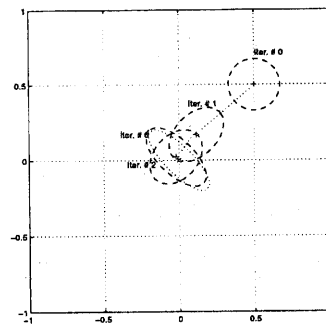


図 5: 問合せ中心の移動

4.4 実データによる実験

実データとして, Montgomery County Dataset [FK94] を用いた. このデータは, アメリカメリーランド州のモンゴメリー郡における道路の交差点を収めたものである (図 6). データ集合は $[-1, 1] \times [-1, 1]$ のに正規化してある. 実験では図の中の正方形で囲んだ部分を使用する.

ここでの問合せは, 高速道路 I-270 (地図で拡大した部分) に沿った点を探すものである. まず, 図 7 で $\mathbf{x}'$ と表している 5 個の点を例として与えた. 繰返しをまったく

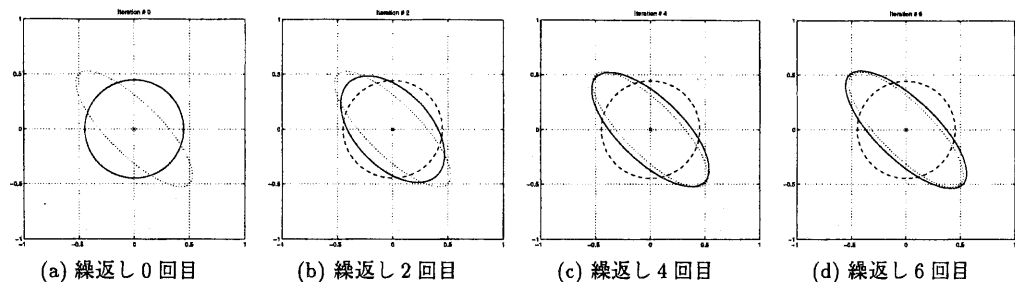


図 4: 繰返しによる問合せの洗練

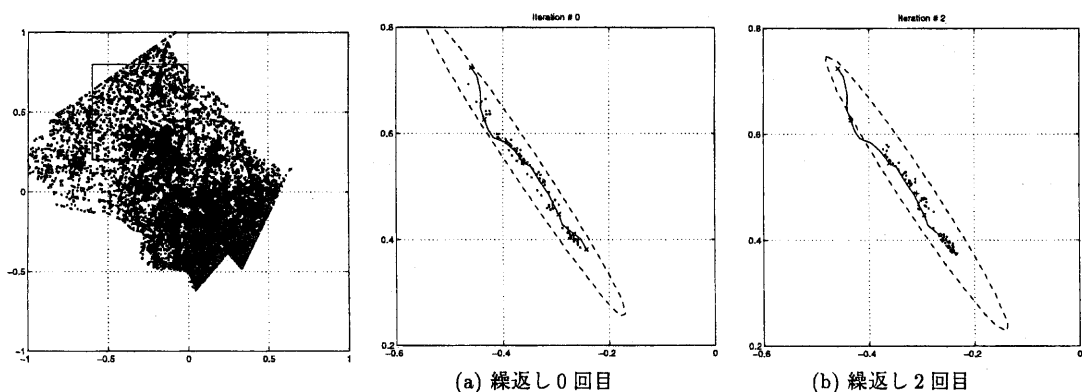


図 6: Montgomery Country Dataset

図 7: 道路に沿った点の問合せ: (a) 与えた 5 個の例 ('x' で表す) と、推定された距離関数の等距離楕円 (繰返しなしの場合); (b) 2 回繰り返した後の等距離楕円

行わなくとも、図 7(a)に見られるように、推定された距離関数の等距離楕円はほとんどの要求される点を含んでいることがわかる。繰返しを行った場合については、図 7(b)に 2 回の繰返し後の等距離楕円を示す。

5 実装における課題

5.1 ユーザインタフェース

本論文で示した MindReader の手法を用いて、ユーザフレンドリーな問合せインタフェースを作成することが可能となる: ユーザはまずブラウジングにより検索をはじめ。ユーザは自身が検索したいデータに類似したものをクリックし、例としてシステムに与える。たとえば画像データベースであれば、サンプル画像が例となる。複数レベルのスコアは、データの選択時にユーザに複数回のクリックをさせることで容易に導入できる。単純には、 i 番目のデータに対しクリックした回数を、スコア v_i とすればよい。

これまでの議論では、検索とフィードバックの繰返しによる問合せ処理ステップを想定していたが、MindReader

では、必ずしも適合フィードバックを用いる必要はない。図 7 の実験で示したように、ユーザがある程度十分な例を与えれば、MindReader は繰返しなしでも的確な推定を下すことが可能である。

5.2 速度

先に述べたように、ここではデータは n 次元空間上の点として表現されていると考えている (例: 地図上の都市、マルチメディアアプリケーションにおける特徴ベクトルなど)。一般に、 n 次元の点データは、X-tree [BKK96], SR-tree [KS97], R*-tree [BKSS90] のような空間アクセスメソッド (spatial access method, SAM) により索引づけすることができる。これまでの SAM の研究により、ユークリッド距離のもとでは SAM は効率的に空間問合せ (範囲問合せ、最近隣問合せ) を支援できる。最近では、Seidl と Kriegel により、重みづけしたユークリッド距離にもとづく問合せのみならず、斜交問合せを既存の SAM の手法により支援する手法が提案されている [SK97]。すなわち、ある一般化された楕円距離とそれにもとづく問合せが与えられれば、SAM を用いて効率的に問合せを処

理することができる。本研究で提案した例からの距離関数の推定手法では、問合せ処理時に距離行列 M の値が変更され、問合せに用いる距離関数が動的に変化することになるが、[SK97] の手法により、このような場合でも SAM を活用できることが保証される。

6 おわりに

本稿では、与えられた複数の例データとそれらに対するスコアをもとに、自動的にユーザが意図している距離関数と問合せ点を推定する手法である MindReader について、その手法と実験結果を述べた。MindReader を用いて検索とフィードバックを繰り返すことにより、ユーザは対話的に自身の問合せを洗練していくことが可能である。本研究の枠組みは、古くは情報検索、最近ではマルチメディアデータベースでも研究されている適合フィードバックの概念を含むものであり、既存の研究の一般化としてとらえることができる。

参考文献

- [BKK96] S. Berchtold et al.: "The X-tree: An Index Structure for High-Dimensional Data", in *Proc. of VLDB*, pp. 28-39, Mumbai, India, Sept. 1996.
- [BKSS90] N. Beckmann et al.: "The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles", in *Proc. of ACM SIGMOD*, pp. 322-331, Atlantic City, NJ, May 1990.
- [FBF⁺94] C. Faloutsos et al.: "Efficient and Effective Querying by Image Content", *Journal of Intelligent Information Systems*, 3(3/4):231-262, July 1994.
- [FK94] C. Faloutsos and I. Kamel: "Beyond Uniformity and Independence: Analysis of R-trees Using the Concept of Fractal Dimension", in *Proc. ACM PODS*, pp. 4-13, Minneapolis, MN, May 1994.
- [GV96] G. H. Golub and C. F. Van Loan: *Matrix Computations*, The Johns Hopkins University Press, Baltimore and London, third edition, 1996.
- [Har92] D. Harman: "Relevance Feedback and Other Query Modification Techniques", in W. B. Frakes and R. Baeza-Yates eds., *Information Retrieval: Data Structures & Algorithms*, pp. 241-263, Prentice-Hall, 1992.
- [ISF98] Y. Ishikawa et al.: "MindReader: Querying databases through multiple examples", in *Proc. of VLDB*, New York, Aug. 1998 (to appear).
- [HK92] K. Hirata and T. Kato: "Query by Visual Example - Content based Image Retrieval -", in *Proc. of 3rd Int'l Conf. on Extending Database Technology (EDBT'92)*, Vol. 580 of *LNCIS*, pp. 56-71, Vienna, Austria, Mar. 1992, Springer-Verlag.
- [KS97] N. Katayama and S. Satoh: "The SR-tree: An Index Structure for High-dimensional Nearest Neighbor Queries", in *Proc. of ACM SIGMOD*, pp. 369-380, Tucson, Arizona, May 1997.
- [Mot88] A. Motro: "VAGUE: A User Interface to Relational Databases that Permits Vague Queries", *ACM TOOLS*, 6(3):187-214, July 1988.
- [RHM97] Y. Rui et al.: "Content-based Image Retrieval with Relevance Feedback in MARS", in *Proc. of IEEE Int'l Conf. on Image Processing '97*, Santa Barbara, CA, Oct. 1997.
- [RHM98] Y. Rui et al.: "Human Perception Subjectivity and Relevance Feedback in Multimedia Information Retrieval", in *Proc. of IS&T and SPIE Storage and Retrieval of Image and Video Databases VI*, San Jose, CA, Jan. 1998.
- [Roc71] J. J. Rocchio: "Relevance Feedback in Information Retrieval", in G. Salton ed., *The SMART Retrieval System - Experiments in Automatic Document Processing*, pp. 313-323, Prentice Hall, Englewood Cliffs, N.J., 1971.
- [SK97] T. Seidl and H.-P. Kriegel: "Efficient User-Adaptable Similarity Search in Large Multimedia Databases", in *Proc. of VLDB*, pp. 506-515, Athens, Greece, Aug. 1997.
- [SL96] A. Spink and R. M. Losee: "Feedback in Information Retrieval", in M. E. Williams ed., *Annual Review of Information Science and Technology (ARIST)*, Vol. 31, chapter 2, pp. 33-78, 1996.
- [Vir] "Virage Inc. Home Page", <http://www.virage.com/>.