

多次元データベースの近似検索のための 量子化次元縮小射影に関する研究

大野 真 吾^{†1} 村上 義 明^{†1} 川 口 尚 志^{†1}
岸 川 直 樹^{†1} 篠 原 武^{†1}

多次元データの近似検索は、一般的に R-tree などの階層的空間索引構造を用いて検索を行う。空間索引構造を用いた近似検索では、高次元データを扱う場合、次元の呪いにより検索性能が悪化することが知られている。これを緩和するため、高次元の特徴データに対して低次元へ射影を行う次元縮小法が提案されている。本論文では、新しい次元縮小法として量子化 Simple-Map(QS-Map) を提案する。QS-Map は、従来用いていた次元縮小法である Simple-Map(S-Map) を量子化したものであり、S-Map と同じ情報量でより効果的な次元縮小が行えると期待できる。従来の S-Map と検索性能を比較することでその有効性を検証する。

Similarity Search of Multi-Dimensional Database by Quantization Dimension Reduction Mapping

SHINGO OHNO,^{†1} YOSHIKI MURAKAMI,^{†1}
HISASI KAWAGUTI,^{†1} NAOKI KISIKAWA^{†1}
and TAKESHI SHINOHARA^{†1}

Hierarchical spatial indexing structures such as R-tree are usually employed for similarity searching on multi-dimensional data. The performances of these indexes become worse on high dimensional feature spaces because of the curse of dimensionality. To avoid this effect, dimension reduction mappings are proposed. In this paper, we propose Quantization Simple-Map(QS-Map) as a dimension reduction technique. QS-Map is quantized S-Map that is a projection method and it is expected to project objects more effectively in the same information volume. We compare QS-Map with usual S-Map to validate its efficiency.

1. はじめに

近似検索とは、蓄積されたデータの中から質問データと似ているデータを探索することであり、パターン認識や機械学習などの様々な分野で役立てることができる。最近では、計算機の高性能化により、画像や音楽等の多数の特徴量を持つ多次元データを大量に扱えるようになり、それに伴い大規模な多次元データベースからの高速な近似検索技術の需要が高まってくると考えられる。多次元データベースに対する近似検索の手法としては、B-tree¹⁾ を多次元に拡張した R-tree²⁾ や M-tree³⁾ 等の階層的な空間索引構造を用いたものがある。これらの索引構造の特徴は、空間的に近いデータ同士を同じ包括領域で囲いノードに登録することである。検索の際には、質問点の検索範囲と重複するノードのみを訪問することにより、参照コストを削減でき、高速な検索を実現している。このような階層的な空間索引構造は、空間の次元が大きくなり過ぎると次元の呪いと呼ばれる現象により、検索効率が悪化してしまうことが知られている。そのため、高次元のデータに対して索引付けを行う場合、次元縮小射影を用いて、データを低次元空間に射影することで、次元の呪いを緩和する必要がある。次元縮小射影の一種として Simple-Map(S-Map)⁴⁾ がある。これは中心点となるデータを選択し、中心点とデータとの距離を射影関数として用いる。S-Map は、任意の距離空間に適用が可能であり、S-Map で低次元に縮小したデータに対し R-tree を構築することで高速な検索を実現できる。

本論文では、新しい次元縮小射影として量子化 Simple-Map(QS-Map) を提案する。QS-Map は、S-Map で用いる中心点とデータとの距離を量子化し、その値を射影関数とする。また、QS-Map の背景には、データの類似性ある程度保持するようにバイナリ文字列で表現する球面分割 (BP) とよばれる手法がある。BP は中心点と半径を設定し、空間を二つの部分空間に分割し射影を行う。また、QS-Map は BP では二つだった分割をより多くの部分空間に分割し、射影を行うとみなすことができる。以上のことから QS-Map は S-Map と BP の一般形であるといえる。QS-Map は、S-Map が保持している値を量子化したものを用いるので、同じ情報量で S-Map より多くの中心点がとることができる。そのため、よりよい射影が行えると考えられ、近似検索の高速化が期待できる。提案した QS-Map で射影したデータを用いて構築した R-tree と従来手法である S-Map で射影したデータを用いて構築した R-tree との検索の比較実験を行う。

^{†1} 九州工業大学院情報工学府情報科学専攻

本論文の構成は以下の通りである．第2章では，近似検索，近似検索を行うための階層的な空間索引構造について紹介する．次元縮小射影 S-Map と提案手法の前身である球面分割 (BP) については第3章で示す．第4章では検索を高速化できる次元縮小射影の提案を行う．提案した次元縮小射影を第5章で検証し，第6章でまとめる．

2. 近似検索システム

2.1 距離空間

近似検索は，質問データからの非近似度により特徴空間内からデータを取得するものである．近似検索システムは，質問点との非近似度が小さいデータを似ていると判断し，それらの特徴空間内から取り出すシステムのことである．我々はデータ間の非近似度として，距離を用いる．

データベースが対象とする特徴空間全体を $\mathcal{U} = \mathbb{R}^n$ とする．ここで， $\mathbb{R}$ は実数全体， n は特徴の次元数を表す．索引の対象となる m 個のデータの集合を $S = \{O_1, O_2, \dots, O_i, \dots, O_m\} \subseteq \mathcal{U}$ とする． O_i は n 次元の特徴 $\{O_i^{(1)}, O_i^{(2)}, \dots, O_i^{(t)}, \dots, O_i^{(n)}\}$ を持つ．ここで， $O_i^{(t)} \in \mathbb{R}$ である．任意の2点のデータ間の非近似度を示す距離関数を $d: \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}^+$ とし， $\mathcal{D} = (\mathcal{U}, d)$ を距離空間とする．距離関数 d は距離の公理と呼ばれる以下の条件を満たすものとする．

- (1) $d(X, Y) \geq 0$ (非負性)
- (2) $d(X, Y) = d(Y, X)$ (対称性)
- (3) $d(X, Y) \leq d(X, Z) + d(Z, Y)$ (三角不等式)
- (4) $d(X, Y) = 0 \Leftrightarrow X = Y$ (同一性)

ここで， $X, Y, Z \in \mathcal{U}$ である．上の条件の中で最も重要なものは，三角不等式である．近似検索システムは，与えられた質問点 $Q = \{Q^{(1)}, Q^{(2)}, \dots, Q^{(t)}, \dots, Q^{(n)}\} \in \mathcal{U}$ に対して近似度が高い（距離が近い）データを結果として与えるものである．

2.2 質問方法

近似検索に用いる質問には，範囲質問や k 近傍質問と呼ばれるものがある．質問点 Q と半径 $r \in \mathbb{R}^+$ を質問のパラメータとする範囲質問 $\text{Range}(\mathcal{D}, Q, r)$ は， Q から距離 r 以内のデータを取得する質問である．すなわち，

$$\text{Range}(\mathcal{D}, Q, r) = \{O_i \in S \mid d(O_i, Q) \leq r\}$$

である．

また，質問点 Q から距離が小さい k 個のデータを取得する質問 $k\text{-NN}(\mathcal{D}, Q, k)$ があり，

これを k 近傍質問という．その解の集合は

$$k\text{-NN}(\mathcal{D}, Q, k) = \{O_i \in S \mid O_i \text{ は } d(O_i, Q) \text{ が小さいものから } k \text{ 個}\}$$

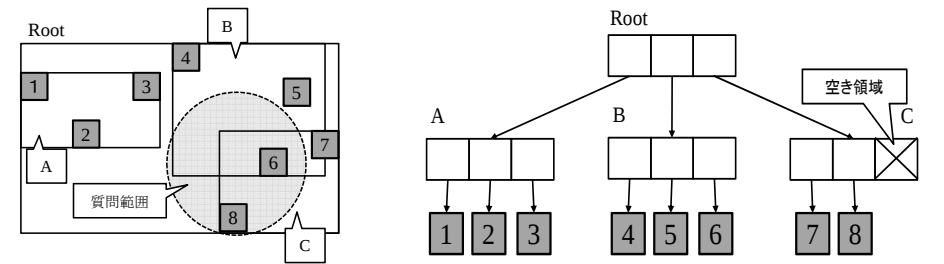
である．特に質問点 Q から距離が最も小さいデータを取得する質問を最近傍質問という．さらに，上記の二つの質問方法を組み合わせた範囲限定 k 近傍質問 $k\text{-NNRange}(\mathcal{D}, Q, r, k)$ がある．これは，質問点 Q から距離が r 以内のものの中で，距離の近い順にデータを k 個取得する質問方法である．つまり，

$$k\text{-NNRange}(\mathcal{D}, Q, r, k) = k\text{-NN}(\mathcal{D}, Q, k) \cap \text{Range}(\mathcal{D}, Q, r)$$

である．

2.3 R-tree

ここでは近似検索に用いられる索引構造 R-tree を紹介する．R-tree は，一次元空間での代表的な索引構造である B-tree を多次元空間に拡張したものである．R-tree はデータを N 次元直方体で囲み，その領域 (MBR: Minimum Bounding Rectangle= 最小包囲矩形) をノードに格納し，索引構造を構築する．各内部ノードはすべての子ノードを包括する MBR と子ノードへのポインタを持っており，葉ノードにはオブジェクト（もしくはオブジェクトへのポインタ）を持っている．R-tree を構築する際には，空間的に近いオブジェクト同士ができるだけ同じノード内に格納されるように MBR を構成する．図1に2次元空間における特徴空間 (a) と，それに対する R-tree の例 (b) (最大要素数 M は3) を示す．図中の1~8はオブジェクトを，A~C は MBR を表す．



(a) 特徴空間 (2次元)

(b) 左図の特徴空間に対する R-tree

図1 R-tree の例 (2次元空間・最大要素数は3)

R-tree を用いた検索では、検索範囲と重なる MBR をもつノードのみを訪問する．検索範囲と重ならない MBR を持つノードを訪問しないことで高速な検索を実現している．図 1 の例では、B と C は検索範囲と重なっているが、A は重なっていない．よって、MBR・A を持つノードは訪問せず、検索範囲内に入っていないオブジェクト 1, 2, 3 への参照を回避することができる．

2.4 射影空間上での R-tree

R-tree を用いた検索では、特徴空間の次元が高次元になると、次元の呪いと呼ばれる現象により検索効率が悪化してしまう．そのため、高次元のデータに対して R-tree を用いる場合、次元縮小射影を用いてデータの次元を下げてから索引構造を構築するのが望ましい．次元縮小射影によって次元が低次元に縮小された空間を射影空間、射影する前の空間を元空間と呼ぶ．

オブジェクト X, Y の n' 次元の座標系への射影をそれぞれ X', Y' とする．ただし、射影上の距離 $d'(X', Y')$ は、以下の式を満たす必要がある．

$$d'(X', Y') \leq d(X, Y)$$

すなわち、射影空間で遠いものは必ず元空間でも遠く、元空間で近いものは必ず射影空間でも近くなるようにしなければならない．元空間で遠いものが、射影空間上で近くなることを、距離が縮むという．

検索では、構築した索引構造に対し、射影空間上でノードを辿っていき、解となるオブジェクトを探索する．解候補となるオブジェクトを見つけると、元空間上の距離を計算することで、真に近いオブジェクトを見つけ出す．

また、検索を高速化するために、解候補となるオブジェクトを見つけた際に元空間上の距離を計算する前に、射影空間上の距離を計算する．射影空間上の距離を計算することで、元空間での距離の下限を見積り、不必要な元空間での距離計算を削減できる．射影距離計算が元空間の距離計算より低コストであるならば、この方法は検索効率を上げるのに有効である．射影距離計算で見積もる距離の下限が大きいほど、元空間での距離計算を多く削減できる．そのため、射影空間上の R-tree で用いる次元縮小射影は、より大きく距離の下限を見積もれるものが望ましい．

3. 次元縮小射影と球面分割 Sketch

本章では、次元縮小射影である Simple-Map と提案手法の前身である球面分割 (BP) に

ついて紹介する．

3.1 Simple-Map

Simple-Map(S-Map) は、中心点とよばれるデータとそれ以外のデータとの距離を座標値として射影する方法である．中心点 $p \in \mathcal{U}$ を用いた、データ X の射影空間上での座標値 $\varphi_p(X)$ を

$$\varphi_p(X) = d(p, X)$$

とする．このとき、三角不等式より $d(X, Y) \geq |\varphi_p(X) - \varphi_p(Y)|$ を満たす．そのため、射影距離は、元の距離に比べ縮むことはあっても伸びることはなく、距離の下限を見積もることができる．次に、中心点を複数用いた場合に拡張する．点の集合を

$$P = \{p_1, p_2, \dots, p_{n'}\} \subset \mathcal{U}$$

と定義する．このとき射影関数を、

$$\varphi_P(X) = (\varphi_{p_1}(X), \varphi_{p_2}(X), \dots, \varphi_{p_{n'}}(X))$$

とする．射影空間上での 2 点間のデータの距離は、

$$d(X, Y) \geq d'(\varphi_P(X), \varphi_P(Y)) = \max_{p \in P} |\varphi_p(X) - \varphi_p(Y)|$$

となる．つまり、射影空間は L_∞ 距離空間となる．中心点の個数を多くすることで、射影空間上での距離の下限をより大きく見積もることが可能である．しかし、中心点の個数を増やすことで射影距離計算のコストが増大し、元空間の距離計算の回数を減らしても検索時間自体を増加してしまう可能性がある．そのため、適切な射影次元を選択する必要がある．

3.2 中心点選択法

S-Map で射影を行う際に、中心点によっては検索効率が悪化してしまうため、どの中心点を選択するかが重要となってくる．まず、データベースから中心点候補となるデータをいくつか取得し、これらを試験的に射影する．この射影データに対し評価関数を適用し、スコア付けを行う．最もスコアが高いものを中心点として採用する．この評価関数に関しては、射影距離総和法を用いる．射影距離総和法は、射影空間上のオブジェクト間の距離の和が最大になる中心点を選択する．

3.3 Sketch

R-tree 以外の検索方法として、Sketch を用いた検索^{5),6)}がある．Sketch を用いた検索は、検索の精度を犠牲にしつつ、検索を高速化する近似アルゴリズム^{7)–10)}の一種である．Sketch とは、実空間での類似性のある程度保持するようにオブジェクトをバイナリ文字列で

表現したものである．Sketch は基礎分割関数を用いて作成される．任意の Sketch 間の距離はハミング距離を用いて計測する．Sketch はバイナリ文字列で表現されているため，Sketch 間の距離はビット演算で高速に計算できる．Sketch を用いた k 近傍検索は 2 段階で行う．

- (1) 質問データから作成した Sketch とデータベース内に登録されているデータの Sketch との距離を計算し，実距離計算する候補データを取り出す．
- (2) 取り出した候補データと質問データとの実空間上の距離を計算する．

Sketch は，実空間上の類似性を完全に保持できるわけではないため，Sketch を用いた検索では精度が悪化する．したがって，より良い類似性を保持した Sketch を作成するためには基礎分割関数の選択が重要である．

3.4 球面分割 Sketch(BPS)

基礎分割関数のひとつである球面分割 (BP)¹¹⁾ は 1 個の中心点と半径を用いた空間分割手法である．BP で作成した Sketch を球面分割 Sketch(BPS) と呼ぶ．任意の距離空間において，集合 $S \in \mathcal{U}$ と中心点 $(p_0) \in \mathcal{U}$ が与えられたとすると，オブジェクト $o \in S$ は以下のように部分空間 S_0, S_1 に分類される．

$$S_0 = \{o \in S | d(p_0, o) \leq r\}$$

$$S_1 = \{o \in S | d(p_0, o) > r\}$$

図 2 は，ユークリッド平面において中心点 P と半径 r を用いて集合 $S = \{A, B, C, D, E, F, G, H, I, J\}$ を分割した例である．

図 2 では，中心点 $P = E$ と半径 r により，集合 S は， $S_0 = \{C, E, F, G, H\}$ と $S_1 = \{A, B, D, I, J\}$ に分割される．ユークリッド平面における BP では，球面を用いてを二分割する．BP を t 回繰り返すことで長さ t の Sketch を作成することができる．Sketch の各ビット $\sigma_i(x)$ は中心点 p と半径 r を用いて

$$\sigma_i(x) = \begin{cases} 0, & \text{if } d(p_i, X) \leq r_i \\ 1, & \text{if } d(p_i, X) > r_i \end{cases} \quad (i = 1, 2, \dots, t) \quad (1)$$

となる．

Sketch は，ハミング距離によって距離が定義される．この場合，距離の下限を見積もることとはできない．しかし，作成された Sketch データと質問点との距離を用いて，以下の式のようにすると

$$\max \begin{cases} |\sigma_i(X) - r_i|, & (\sigma_i(X) \neq \sigma_i(Y)) \\ 0, & (\sigma_i(X) = \sigma_i(Y)) \end{cases} \quad (i = 1, 2, \dots, t) \quad (2)$$

となり距離の下限を見積もることができる^{12),13)}．これは三角不等式を用いることで以下の式で証明ができる．

$(\sigma_i(X) \neq \sigma_i(Y))$ の場合

$$\begin{aligned} d(Q, O) &\geq |d(Q, p) - d(O, p)| \\ &\geq |d(Q, p) - r_i| \end{aligned}$$

$(\sigma_i(X) = \sigma_i(Y))$ の場合

$$d(Q, O) \geq |d(Q, p) - d(O, p)| \geq 0$$

となる．

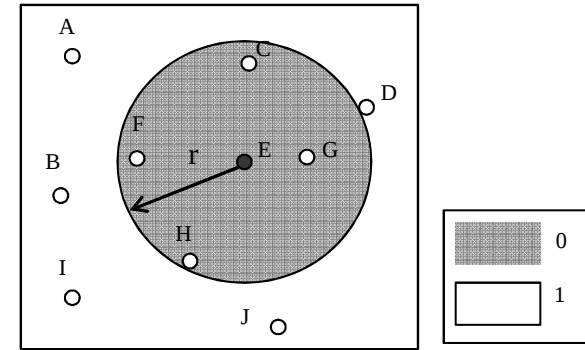


図 2 ユークリッド平面における BP

4. 量子化 Simple-Map(QS-Map)

本章では提案手法である量子化 Simple-Map(QS-Map) について紹介する．

BP は半径を用いて空間を二分割し射影している．次に BP を用いて，空間を二つ以上に分割するように拡張する．点 $v \in \mathcal{U}$ ，刻み幅 ρ と分割数 m ($\rho, m \in \mathbb{R}^+$) を用いた，データ X の射影空間上での座標値 $\sigma_v(X)$ を，

$$\sigma_v(X) = \begin{cases} \left\lfloor \frac{d(v, X)}{\rho} \right\rfloor & \left(\frac{d(v, X)}{\rho} < m-1 \right) \\ m-1 & \left(\frac{d(v, X)}{\rho} \geq m-1 \right) \end{cases}$$

となる．以上のような射影方法を量子化 Simple-Map(QS-Map) と呼ぶ．図 3 は，ユークリッド平面において中心点 P と刻み幅 ρ ，分割数を 4 にして，集合 $S = \{A, B, C, D\}$ を射影した例である．

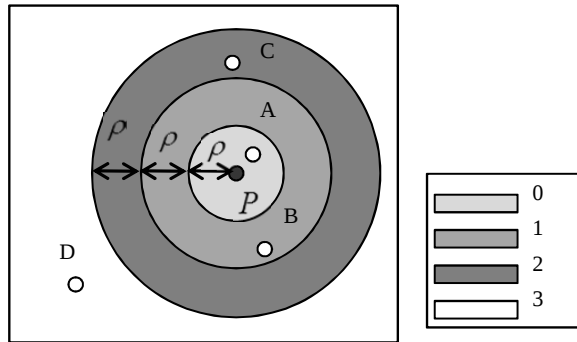


図 3 ユークリッド平面における QS-Map

図 3 では，中心点 P と刻み幅 ρ により， $\sigma_P(A) = 0$ ， $\sigma_P(B) = 1$ ， $\sigma_P(C) = 2$ ， $\sigma_P(D) = 3$ となる．QS-Map は，中心点とその他のデータ間の距離を量子化した値に射影する．また，刻み幅 $\rho = 1$ と分割数 $m = \infty$ の場合を考えると S-Map と見なすことができる．つまり，QS-Map は BP と S-Map の一般的な形といえる．

次に射影距離について紹介する．質問点 Q とオブジェクト O の間の QS-Map での射影距離は

$$\begin{cases} \rho(|\sigma_v(Q) - \sigma_v(O)| - 1) & (\sigma_v(Q) \neq \sigma_v(O)) \\ 0 & (\sigma_v(Q) = \sigma_v(O)) \end{cases}$$

上の式は，実距離に比べて伸びることはなく，距離の下限として使用できる．これは以下の式で証明できる ($a = \varphi_v(Q) - \rho\sigma_v(Q)$, $b = \varphi_v(O) - \rho\sigma_v(O)$ とする)．

- (1) $\sigma_v(Q) = \sigma_v(O)$ のとき
 $d(Q, O) \geq 0$
- (2) $\sigma_v(Q) \neq \sigma_v(O)$ のとき
 $d(Q, O) \geq (\varphi_v(Q) - \varphi_v(O))$
 $= |(\rho\sigma_v(Q) + a) - (\rho\sigma_v(O) + b)|$
 $= |(\rho\sigma_v(Q) - \rho\sigma_v(O)) + (a - b)|$
 $\geq |\rho\sigma_v(Q) - \rho\sigma_v(O)| - |a - b|$
 $\geq \rho(|\sigma_v(Q) - \sigma_v(O)| - 1)$

多くの場合，QS-Map の射影距離は S-Map の射影距離に比べ，最大 2ρ の離散化誤差が生じる．また， O はデータベースに格納するとき QS-Map に射影するが， Q は検索時に直接，中心点との距離を計算し射影する．そのため，質問点 Q は，S-Map の射影が使うことができ，距離の下限をより大きく見積もることができる (図 4)．これにより，質問点を射影する際の QS-Map の離散化誤差を 2ρ から ρ に軽減することができる．オブジェクト間の射影距離は，S-Map の射影と QS-Map の射影を用いることにより以下の式で表すことができる．これは Sketch での距離の下限を見積もる (2) 式に対応している^{12),13)}

$$\begin{cases} \varphi_v(Q) - \rho(\sigma_v(O) + 1) & (\sigma_v(Q) > \sigma_v(O)) \\ \rho\sigma_v(O) - \varphi_v(Q) & (\sigma_v(Q) < \sigma_v(O)) \\ 0 & (\sigma_v(Q) = \sigma_v(O)) \end{cases}$$

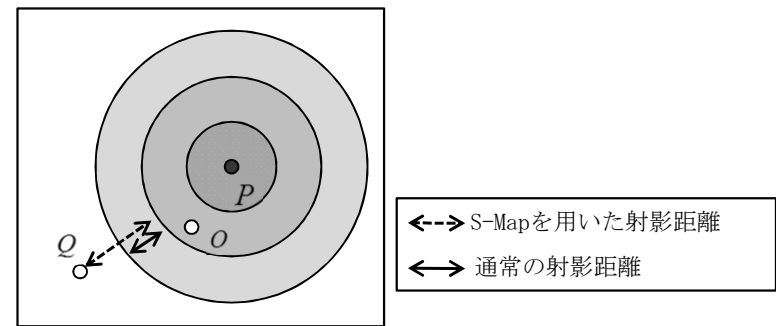


図 4 S-Map 射影を用いた射影距離

この射影距離は実距離より伸びることはないことを以下の式で示す．

(1) $\sigma_v(Q) > \sigma_v(O)$ のとき

$$\begin{aligned} d(Q, O) &\geq \varphi_v(Q) - \varphi_v(O) \\ &= \varphi_v(Q) - (\rho\sigma_v(O) + b) \\ &\geq \varphi_v(Q) - \rho(\sigma_v(O) + 1) \end{aligned}$$

(2) $\sigma_v(Q) < \sigma_v(O)$ のとき

$$\begin{aligned} d(Q, O) &\geq (\varphi_v(O) - \varphi_v(Q)) \\ &= \rho\sigma_v(O) + b - \varphi_v(Q) \\ &\geq \rho\sigma_v(O) - \varphi_v(Q) \end{aligned}$$

(3) $\sigma_v(Q) = \sigma_v(O)$ のとき

$$d(Q, O) \geq |\varphi_v(Q) - \varphi_v(O)| \geq 0$$

次に、中心点を複数用いた場合に拡張する．点の集合を $V = \{v_1, v_2, \dots, v_{n'}\} \subset \mathcal{U}$ と定義する．このとき射影関数を、

$$\sigma_V(X) = (\sigma_{v_1}(X), \sigma_{v_2}(X), \dots, \sigma_{v_{n'}}(X))$$

とする．射影空間上での 2 点間のデータの距離は、

$$\max_{v \in V} \begin{cases} d(Q, O) \geq d''(\varphi_V(Q), \sigma_V(O)) = \\ \varphi_v(Q) - \rho(\sigma_v(O) + 1) & (\sigma_v(Q) > \sigma_v(O)) \\ \rho\sigma_v(O) - \varphi_v(Q) & (\sigma_v(Q) < \sigma_v(O)) \\ 0 & (\sigma_v(Q) = \sigma_v(O)) \end{cases}$$

となる．

QS-Map は、前章で紹介した次元縮小射影 SMap と BP の一般的な形とみなすことができる．QS-Map は S-Map に比べて以下の二つの利点があると考えられる．

(1) 距離の下限をより大きく見積もることが可能

QS-Map は、中心点とオブジェクト間の距離を量子化した値を射影する．こうすることで S-Map と同じ情報量でより多くの中心点が取れ、射影距離を大きく見積もることができると考えられる．

(2) 射影距離計算コストの削減

QS-Map の射影距離計算は、分割数分だけ配列を用意し、質問点との距離の下限をあらかじめすべて表に書き出す．射影距離を測る際には、対応する表を呼び出すだけ

でよく射影距離計算のコストを削減できる．

4.1 中心点選択

QS-Map も中心点を用いた射影法であるため、S-Map と同様に中心点選択が重要である．基本的な中心点選択の流れは S-Map と同様である．評価関数である射影距離総和法に関して 2 種類の手法を用いて実験を行った．

(1) 射影距離を S-Map 射影で計測したもの

(2) 射影距離を QS-Map 射影で計測したもの

(1) は S-Map と同様の中心点選択を行う．(2) は射影距離を QS-Map の射影を用いて計測する．また、QS-Map は刻み幅 ρ を設定する必要がある． ρ は中心点を選び終わった後で、データベースと中心点の距離が最大のものを分割数で割ったものの採用する．

5. 実験

S-Map を用いて R-tree を作成した場合と QS-Map を用いて R-tree を作成した場合について性能の比較する．S-Map は 8 次元に射影を行い、各次元は 16bit の整数型で表現する．QS-Map は分割数を 256 とすることで、各次元は 8bit の整数で表現することでき、射影データの情報量が同じになるように 16 次元に射影する．QS-Map の中心点選択では 4.1 節でも述べたように、QS-Map(1) (射影距離を S-Map 射影で計測したもの) と QS-Map(2) (射影距離を QS-Map 射影で計測したもの) の二種類を用いる．訪問ノード数、距離計算回数、検索時間に関して、平均と最近傍解の距離との関係を比較する．使用するデータベースとして画像データと音楽データを用いる．

5.1 画像データ

画像データは、動画から特徴抽出した 64 次元のデータ約 700 万件をデータベースとして用いる．質問データには、データベース内のデータと非常に距離が近いデータを多く持つ質問 (近質問)、データベース内のデータと比較的距離が近いデータを持つ質問 (準近質問)、データベース内に似ているデータが存在しない質問 (遠質問) を各々約 3 万件ずつ用意した．実際の最近傍解との距離の分布を図 5 に示す．平均の検索時間、訪問ノード数、距離計算回数との関係を比較する (表 1)．

5.2 音楽データ

音楽データは、約 700 万件の音楽フレームから特徴抽出した 96 次元のデータをデータベースとして用いる．質問データには、画像データと同様にデータベース内のデータと非常

に距離が近いデータを多く持つ質問 (近質問), データベース内のデータと比較的距離に近いデータを持つ質問 (準近質問), データベース内に似ているデータが存在しない質問 (遠質問) を各々約 5 万件ずつ用意した. 実際の最近傍との距離の分布を図 6 に示す. 平均の訪問ノード数, 距離計算回数, 検索時間との関係进行比较する (表 2).

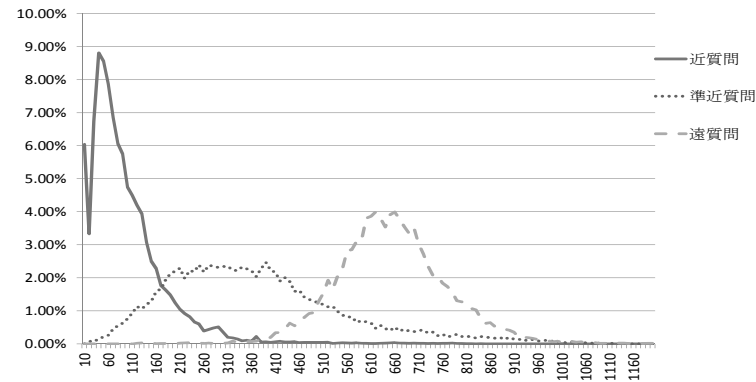


図 5 画像データの最近傍との距離

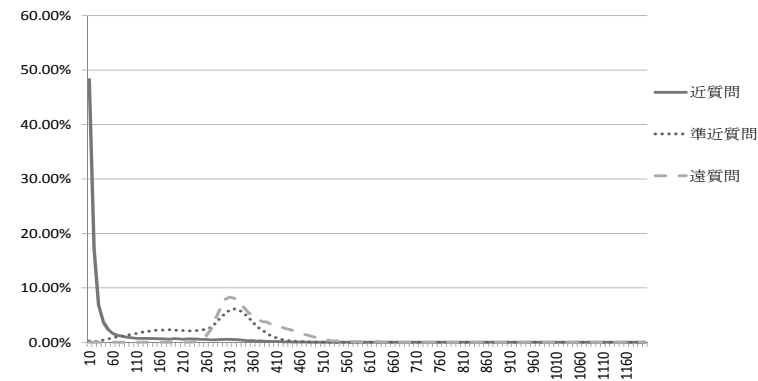


図 6 音楽データの最近傍との距離

表 1 画像データに対する S-Map と QS-Map との比較

	S-Map	QS-Map(1)	QS-Map(2)
訪問ノード数	1.30 万回	1.37 万回	1.40 万回
距離計算回数	68.8 万回	45.6 万回	45.6 万回
検索時間	84.4msec	76.6msec	80.0msec

表 2 音楽データに対する S-Map と QS-Map との比較

	S-Map	QS-Map(1)	QS-Map(2)
訪問ノード数	0.97 万回	0.99 万回	1.00 万回
距離計算回数	65.4 万回	56.5 万回	60.0 万回
検索時間	91.1msec	88.3msec	89.6msec

6. ま と め

実験結果から, 画像データに関しては, QS-Map(1) (射影距離を S-Map 射影で計測したもの) が最も検索時間が早く, S-Map を用いた R-tree に比べ訪問ノード数を 2 % 増加したが, 距離計算回数を 30 % 程, 検索時間を 10 % 程削減することに成功した. 音楽データに関しては, これも QS-Map(1) が性能がよくなっている. S-Map を用いた R-tree に比べ, 訪問ノード数を 5 % 増加したが, 距離計算回数を 10 % 程, 検索時間を 5 % を程削減することに成功した. QS-Map は S-Map に比べ優れた次元縮小射影であるといえる. 訪問ノード数が増加した原因としては, ヒルベルト曲線順序付けによるクラスタリングの相性や MBR の形状が大きくなってしまったことが考えられる. そのため, この観点からの改善が必要である.

QS-Map の中心点選択では, S-Map での射影距離を用いた QS-Map の方が, QS-Map での射影距離を用いた QS-Map より良くなっている. これは局所探索する際に, QS-Map での射影距離を用いた中心点選択では, 評価関数をうまく更新できていないことが考えられる. より良い中心点選択法については今後の課題とする.

最後に, 今回は分割数を固定して実験していたため, 適切な分割数の選択法の考察や, 動画同定システムや楽曲同定システムのような応用研究に今回用いた手法を用いて有効性を検証していく必要がある.

参 考 文 献

- 1) R.Bayer and E.McCreight. Organization and maintenance of large ordered indexes. *Acta Informatica*, Vol.1, pp. 173–189, September 1972.
- 2) A.Guttman. R-trees: A dynamic index structure for spatial searching. In Beatrice Yormark, editor, *SIGMOD'84, Proceedings of Annual Meeting, Boston, Massachusetts, June 18-21, 1984*, pp. 47–57. ACM Press, 1984.
- 3) P.Ciaccia, M.Patella, P.Zezula. M-tree: An efficient access method for similarity search in metric spaces. *VLBD '97 Proceedings of the 23rd International Conference on Very Large Data Bases*, pp. 426–435, 1997.
- 4) T.Shinohara and H.Ishizaka. On dimension reduction mappings for approximate retrieval of multi-dimensional data. *Lecture Notes in Computer Science*, Vol. 2281, pp. 89–94, 2002.
- 5) Q.Lv, M.Charikar, and K.Li. Image similarity search with compact data structures. *Conference on Information and Knowledge Management archive, Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pp. 208–217, 2004.
- 6) Q.Lv, W.Josephson, Z.Wang, M.Charikar, and K.Li. Efficient filtering with sketches in the ferret toolkit. *International Multimedia Conference archive, Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pp. 279–288, 2006.
- 7) G.Amato. *Approximate Similarity Search in Metric Spaces*. PhD thesis, University of Dartmund, 2002.
- 8) G.Amato, F.Rabitti, P.Savino, and P.Zezula. Region proximity in metric spaces and its use for approximate similarity search. *AACM Transactions on Information Systems*, 2003.
- 9) P.Zezula, G.Amato, V.Dohnal, and M.Batko. *Similarity Search: The Metric Space Approach*. Advances in Database Systems. Springer Verlag, 2006.
- 10) P.Zezula, P.Savino, G.Amato, and F.Rabitti. Approximate similarity retrieval with m-trees. *The VLDB Journal - The International Journal on Very Large Data Bases*, 1998.
- 11) 大野真吾, 岩崎瑶平, 篠原武. 空間検索のための Sketch の基礎分割関数に関する研究. 火の国情報シンポジウム 2010, 2010.
- 12) 岩崎瑶平, 篠原武. 局所性保持ハッシングを用いた空間検索法の高速化に関する研究. 電機関係学会第 63 回九州支部連合大会 (第 63 回連合大会), 2010.
- 13) 岩崎瑶平. 座標分割法を用いた局所性鋭敏ハッシング (lsh) による近似検索の高性能化に関する研究. 九州工業大学大学院 修士論文, 2011.