最近接点の有意性の評価によるマルチメディア情報の効率的な検索法

片山紀生　佐藤真一

国立情報学研究所
〒 101-8430 東京都千代田区一ツ橋 2-1-2
TEL (03) 4212-2620　E-mail {katayama,satoh}@nii.ac.jp

キーワード：最近接点の有意性、マルチメディア情報、類似検索、最近接点探索

高次元空間での最近接点探索は、マルチメディア情報を類似検索する手段として広く使われている。ところが、最近の研究成果から、高次元空間では、点相互の距離に有意な差が生じないことがあり、その場合、最近接点の意味が小さくなってしまうことが明らかになっている。このような最近接点は利用者にとって意味が小さいだけでなく、最近接点探索の処理効率を下げる原因にもなる。そこで、我々は、最近接点の有意性を評価する手法を考案するとともに、新しい最近接点探索法として、「有意性感応型最近接点探索 (significance-sensitive nearest neighbor search)」を考案した。この探索法は、最近接点の有意性を評価できるだけでなく、探索コストの低減も可能にする。

Efficient Retrieval of Multimedia Information with Estimating the Significance of Nearest Neighbors

Norio Katayama　Shin'ichi Satoh

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo 101-8430, Japan
TEL +81-3-4212-2620　E-mail {katayama,satoh}@nii.ac.jp

Keyword : Significance of Nearest Neighbors, Multimedia Information, Similarity Retrieval, Nearest Neighbor Search

Nearest-neighbor (NN) search in high-dimensional space is widely used for the similarity retrieval of multimedia information. Recent research results in the literature reveal that NN-search might return insignificant NNs in high-dimensional space. Insignificant NNs are troublesome with respect to the efficiency of the similarity retrieval. Hence, we devised a way to estimate the significance of NNs based on the local intrinsic dimensionality. Then, with applying it, we developed a new NN-search algorithm: the significance-sensitive nearest-neighbor search. This algorithm not only enables us to distinguish more significant NNs from less significant ones but also enables us to cut down the search cost compared with the conventional NN-search algorithm.

1　はじめに

高次元空間における最近接点探索は、マルチメディア情報の類似検索手法として広く使われている。個々のマルチメディア情報は、多次元空間中のベクトルにマップされ、ベクトル間の距離（例えば、ベクトルの先端間のユークリッド距離）が、マルチメディア情報の類似度に相当する。このようなベクトルと多次元空間は、それぞれ、特徴ベクトル、特徴空間と呼ばれている。特徴空間の例としては、画像のカラーヒストグラムが挙げられる。特徴空間の重要な性質のひとつは、高次元であることである。例えば、画像のカラーヒストグラムの場合、しばしば16次元以上の特徴ベクトルが使われる。

最近の研究成果から、高次元空間では、低次元空間では想像できないような興味深い現象が起こることが明らかになっている。高次元空間の自由度があまりにも高いために点が広く散在してしまい、点相互の距離に有意な差が生じないことが起こり得るのである。典型的な例は、単位超立方体中に点が一様に分布している場合であり、片山ら[1, 2]は、それらの点の相互の距離にほとんど差がないことを実験的に示している。そして最近、Beyerら[3]によって、一様分布よりも広い条件のもとでも、最近接点までの距離が、最も遠い点までの距離に、次元が高くなるにつれて漸近することが示されている。

高次元空間では距離に有意な差が生じないという現象は、類似検索の処理効率という観点で問題となる。最近接点の意味が小さい場合、検索結果もユーザにとって意味の小さいものになる上、最近接点探索の処理効率も低下する。というのは、同様な類似度を持つ多くの点から最近接点を見つけ出さなければならないからである。

しかし、実世界のアプリケーションでは、データの分布に偏りがあるため、ある領域では有意な最近接点がなかったとしても、他の領域では存在することが期待できる。そのため、もし、有意な最近接点とそうでないものを区別することができれば、類似検索の効率を向上できると考えられる。そこで、我々は、局所的な埋め込み次元に基づいて、最近接点の有意性を評価する手法を考案した。そして、それを応用することによって、新しい最近接点探索法として、「有意性感応型最近接点探索 (significance-sensitive nearest neighbor search)」を考案した。この探索法は、最近接点の有意性を評価できるだけでなく、従来の最近接点探索法に比べて、探索コストの低減も可能にする。

この論文の構成は、以下のとおりである。2節では、有意性の低い最近接点が、マルチメディア情報の類似検索に、どのような影響を及ぼすのか説明する。3節では、最近接点の有意性を評価する手法について、4節では、有意性感応型最近接点探索法について説明する。そして、評価実験の結果を5節で示し、6節で結論を述べる。

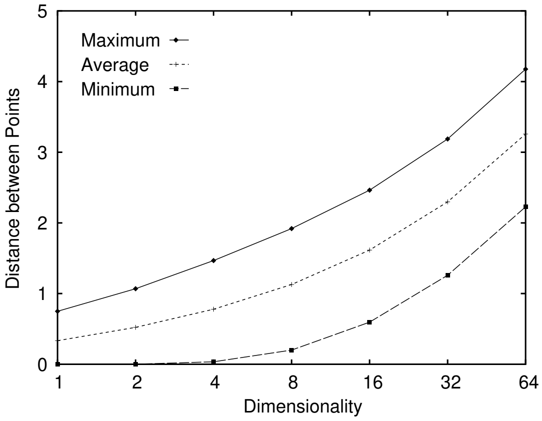


図1: 単位超立方体に一様に生成した100,000個の点の間の距離。([2]から引用).

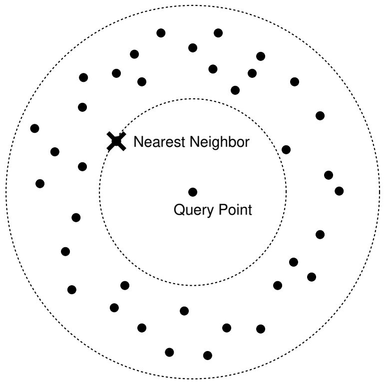


図2: 有意性の低い最近接点。

2　最近接点の有意性と類似検索

2.1　有意性の低い最近接点

最近接点探索を使う場合、我々は、見つかった最近接点が、他の点よりもはるかに近くにあることを期待する。しかし、この直感は、高次元空間では必ずしも成立しない。例えば、点が単位超立方体の中に一様に分布している場合、二点間の距離は、あらゆる組合せについて見ても、ほとんど差が生じないことがある。図1は、100,000個の点を一様に単位超立方体の中に生成し、あらゆる組合せについて距離を求め、その最小値、平均値、最大値を示したものである。図が示すとおり、次元が高くなるにつれて、距離の最小値が著しく増加しており、最大値に対する最小値の比は、16次元で24%、32次元で40%、64次元で53%になっている。したがって、64次元空間では、最近接点までの距離が、最も遠い点までの距離のわずか53%以下になってしまっているのである（図2）。このような場合、我々は、これらの最近接点を、有意性の低い最近接点と見做すことが可能である。なぜならば、最近接点と他の点との差は無視できるほど小さく、他の点は、最近接点と同程度に、質問点（最近接点探索の基準となる点）の近くに存在しているからである。類似検索の視点から言えば、最近接点の有意性が低い

場合、最近接点は他の点とほぼ同程度の類似性を持ち、質問点に対して有意な類似性を持っていないことになる。

図 1が示すとおり、有意性の低い最近接点は、次元が高くなるほど起こりやすくなる。この特性は、k 番めの最近接点までの距離の期待値を求めることによって確認できる。N 個の点が、質問点を中心とする超球の中に一様に分布しているとき、k 番めの最近接点までの距離の期待値 d_{kNN} は、次式のようになる [4]:

$$E\{d_{kNN}\} \approx \frac{\Gamma(k+1/n)}{\Gamma(k)}\frac{\Gamma(N+1)}{\Gamma(N+1+1/n)}r, \tag{1}$$

ここに、n は空間の次元数であり、r は超球の半径である。そして、k 番めの最近接点までの距離と $k+1$ 番めの最近接点までの距離の比は、次式のようになる [4]:

$$\frac{E\{d_{(k+1)NN}\}}{E\{d_{kNN}\}} \approx 1 + \frac{1}{kn}. \tag{2}$$

このように、質問点の周囲に点が一様に分布している場合には、k 番めの最近接点と $k+1$ 番めの最近接点の比は、次元が高くなるほど小さくなることが期待されるのである。これは、有意性の低い最近接点が、低次元空間よりも高次元空間で起こりやすいことを示している。

2.2 有意性の低い最近接点の類似検索への影響

有意性の低い最近接点は、類似検索に対して、以下の点で悪影響を及ぼす。

- 最近接点探索の効率が低下する。

  最近接点の有意性が低い場合、最近接点と同程度の類似性を持つ点が多数存在する。それらの点は、最近接点の有力候補であるため、真の最近接点を特定するまでに、探索処理は、多くの点を調べなければならなくなる。そのため、探索処理の効率が低下する原因となる。

- 意味のない結果が返される。

  最近接点の有意性が低い場合、探索処理の結果は、同程度の類似性を持つ多くの有力候補の中から最も近いものになる。この場合、全ての候補が同程度の類似性を持っているのであるから、全てが類似しているか、全てが類似していないかのどちらかである。したがって、多くの有力候補の中から、真の最近接点を選び出すのは、類似検索としては、あまり意味のないことである。

これらの影響は、人間と対話的に動作する検索システムにおいて特に問題となる。最近接点の有意性が低い場合、検索システムは、意味のない結果を返すまで、ユーザを待たせることになる。そのため、マルチメディア情報の類似検索を効率良く実現するためには、有意性の低い最近接点をどのように扱うかが問題となる。

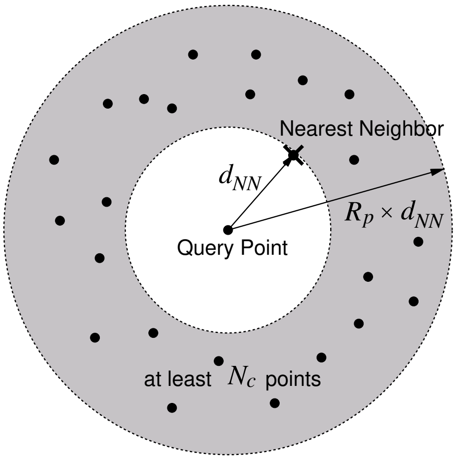


図 3: 有意性の低い最近接点の定義。

3 最近接点の有意性の評価法

3.1 有意性の低い最近接点の定義

既に述べたとおり、有意性の低い最近接点は高次元空間で起こりやすく、類似検索システムにとって厄介な存在となる。しかし、だからと言って、高次元空間が、マルチメディア情報の類似検索に役立たないということではない。実世界のアプリケーションでは、データの分布に偏りがあるため、埋め込み次元数（有効な次元数）は、特徴空間の次元数よりも小さいことが期待できる。例えば、データの分布がいくつかの優勢な次元によって支配されている場合には、埋め込み次元数は、それらの優勢な次元の数になる。さらに、埋め込み次元数は、データセット全体で一貫しているとは限らず、局所的な領域ごとに異なっている可能性がある。したがって、ある領域で最近接点の有意性が低くても、別の領域では最近接点の有意性が高い場合が起こり得るのである。これは、最近接点の有意性を評価できれば、類似検索の効率を高めることが可能であることを示している。そこで我々は、最近接点の有意性を、局所的な埋め込み次元数に基づいて評価する手法を考案した。

我々は、まず、「有意性の低い最近接点」の定義として、次の定義を考案した。

定義 1 d_{NN} を質問点から最近接点までの距離とする。また、質問点からの距離が d_{NN} から $R_p \times d_{NN}$ の範囲にある領域を、質問点の近傍領域と呼ぶことにする。このとき、近傍領域に N_c 個以上の点が存在すれば、その最近接点を「有意性の低い最近接点」と見做す。

ここに、$R_p(>1)$ と $N_c(>1)$ は、制御パラメータである（図 3）。R_p は質問点の近傍領域を決定し、N_c は近傍領域の輻輳を決定する。例えば 5節の実験では、R_p として 1.84 を、N_c として 48 を使っている。R_p と N_c の適切な値を見つける方法は、この節の後半で説明する。

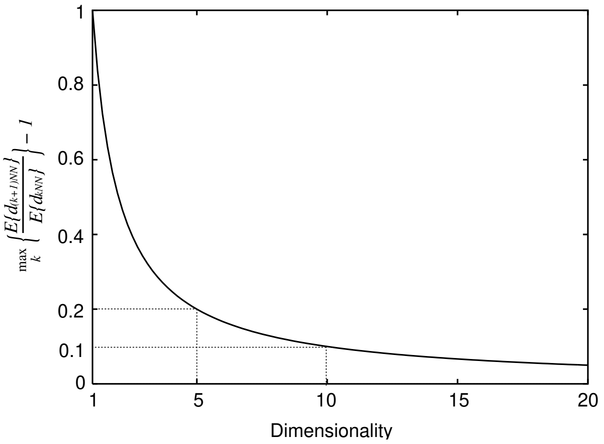


図 4: k 番めと $k+1$ 番めの最近接点の相対差の期待値。

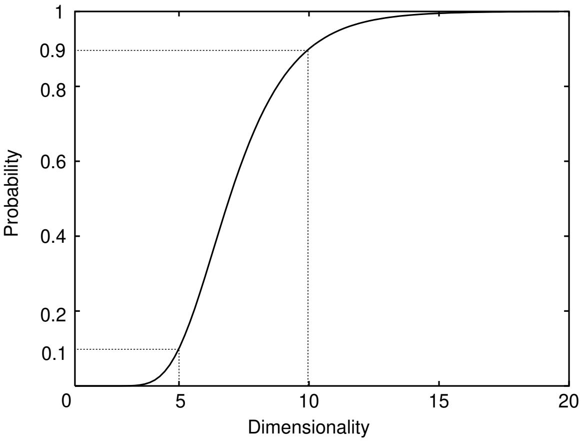


図 5: R_p が 1.84471 N_c が 48.0277 の場合の棄却率。

3.2 埋め込み次元数に基づくパラメータの設定

上記のように、R_p と N_c は、有意性の低い最近接点の定義において、本質的な役割を果たしている。ここで我々は、埋め込み次元に基づいてパラメータを設定する方法を示す。以下に示す設定法は、二段階から成る。まず、前半で、有意性の低い最近接点と埋め込み次元数とを関連づけ、二つの制御点 (ν_c, ρ_c) と (ν_r, ρ_r) を決定する。後述のとおり、ν_c はカットオフ次元数、ν_r は遮断次元数であり、ρ_c は ν_c での棄却率、ρ_r は ν_r での棄却率である。これらの制御点によって、いくつ以上の埋め込み次元数を、最近接点の有意性が低くなる次元数と見做すか決定する。次に、後半で、R_p と N_c を二つの制御点から決定する。

まず最初に、最近接点の有意性と埋め込み次元数とを、式 (2) によって関連付ける。この式は、埋め込み次元数が n のときにも成り立つ [4]。すなわち、この式の n は、データ空間の次元数を意味しているのではなく、有効な次元数（すなわち、埋め込み次元数）を意味しているのである。式 (2) は、$d_{(k+1)NN}$ と d_{kNN} の期待値の比が、k が増加するにつれて単調減少することを示している。したがって、期待値の比が最大となるのは、次式のように 1 番めと 2 番めの最近接点においてである。

$$\max_k \frac{E\{d_{(k+1)NN}\}}{E\{d_{kNN}\}} = \frac{E\{d_{2NN}\}}{E\{d_{1NN}\}} \approx 1 + \frac{1}{n}. \qquad (3)$$

この式は、k 番めの最近接点と $k+1$ 番めの最近接点の相対差の期待値が、次元数が高くなるにつれて単調減少することを示している（図 4）。5 次元の場合には、20% 以下の差しか期待できず、10 次元の場合には、10% 以下の差しか期待できないことがわかる。このように、埋め込み次元数は、最近接点の相対的な有意性を見積もることを可能にする。我々は、以下で述べるように、この性質を使って、二つの制御点 (ν_c, ρ_c) と (ν_r, ρ_r) を決定する。

次に、埋め込み次元数 n とパラメータ R_p、N_c を関連付ける。局所的な領域において、埋め込み次元数 n で、点が一様に分布していると仮定すると、R_p で決定される近傍領域に N_c 個以上の点が存在する確率は、次式のようになる。

$$Prob\{\text{at least } N_c \text{ points in } R_p\} = (1 - (1/R_p)^n)^{N_c} \qquad (4)$$

我々の定義によれば、R_p で決定される近傍領域に N_c 個以上の点が存在すれば、最近接点の有意性が低いと見做されるので、式 (4) は、埋め込み次元数が n のときに、最近接点の有意性が低いと見做される確率を示していることになる。そこで我々は、この確率を「(最近接点の) 棄却率」と呼ぶことにする。棄却率は、埋め込み次元数が大きくなるにつれて単調に増加する。また、以下に示すように、二つの制御点によって、容易に制御することが可能である。今、棄却率を、次元数 ν_1 で ρ_1、次元数 ν_2 で ρ_2 に設定したいとする（$0 < \rho 1 < \rho 2 < 1$ かつ $1 < \nu_1 < \nu_2$）。すると、パラメータ R_p と N_c は、次の連立方程式を解くことによって決定できる。

$$(1 - (1/R_p)^{\nu_1})^{N_c} = \rho_1 \qquad (5)$$

$$(1 - (1/R_p)^{\nu_2})^{N_c} = \rho_2 \qquad (6)$$

N_c を消去することによって次式を得る。

$$\frac{\log(1 - (1/R_p)^{\nu_1})}{\log(1 - (1/R_p)^{\nu_2})} = \frac{\log \rho_1}{\log \rho_2} \qquad (7)$$

この方程式は、算術的に解くことはできない。しかし、左辺が R_p に対して単調に増加するため、ニュートン法などの数値解法によって容易に解くことができる。R_p が求まれば、次式によって N_c も求まる。

$$N_c = \frac{\log \rho_1}{\log(1 - (1/R_p)^{\nu_1})}. \qquad (8)$$

以上のように、式 (7) と (8) を用いることで、二つの制御点 (ν_1, ρ_1) と (ν_2, ρ_2) から、R_p と N_c を決定することが可能になる。

次に問題になるのが、どのようにして制御点を決定するかであるが、我々は、これらの制御点を、低域通過型フィルタとのアナロジーから、カットオフ点 (ν_c, ρ_c) と遮断点

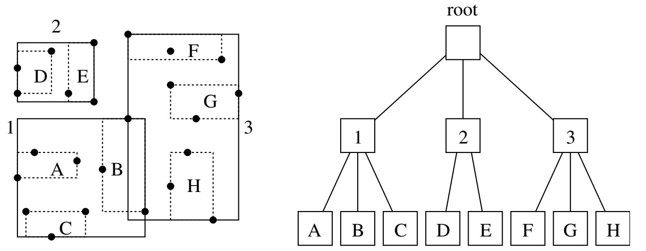


図 6: R-tree の階層構造。

(ν_r, ρ_r) のペアとして設定する方法を提案する。$n < \nu_c$ の範囲は、通過域に相当し、埋め込み次元数がこの範囲にあると、高い確率で有意性が高いと見做される。一方、$n > \nu_r$ の範囲は、遮断域に相当し、埋め込み次元数がこの範囲にあると、高い確率で有意性が低いと見做される。$\nu_c < n < \nu_r$ の範囲は、過渡域に相当する。また、我々は、ν_c と ν_r を k 番めと $k+1$ 番めの最近接点の相対差の近似的な期待値（式 (3) と図 4）によって決定することを提案する。例えば、5 次元では 20% 以下の差しか期待できず、10 次元では 10% 以下の差しか期待できないので、カットオフ次元数を 5、遮断次元数を 10 とすることが考えられる。そして、カットオフ次元数での棄却率を 0.1、遮断次元数での棄却率を 0.9 とすると、二つの制御点は、(5, 0.1) と (10, 0.9) となる。これらの制御点から、式 (7) と (8) によって、R_p と N_c を求めると、R_p が 1.84471、N_c が 48.0277 となる。図 5 は、これらのパラメータを用いたときの棄却率を、式 (4) によって計算した結果である。

以上のように、ここで提案した手法を用いることで、最近接点の相対差と棄却率に基づいて、パラメータ R_p と N_c を決定することが可能になる。すなわち、期待される相対差から、最近接点の有意性を高いとする次元数の範囲と、低いとする次元数の範囲を定め、制御点として、低域通過型フィルタのように、カットオフ点と遮断点を決めることで、パラメータの値が決まるのである。

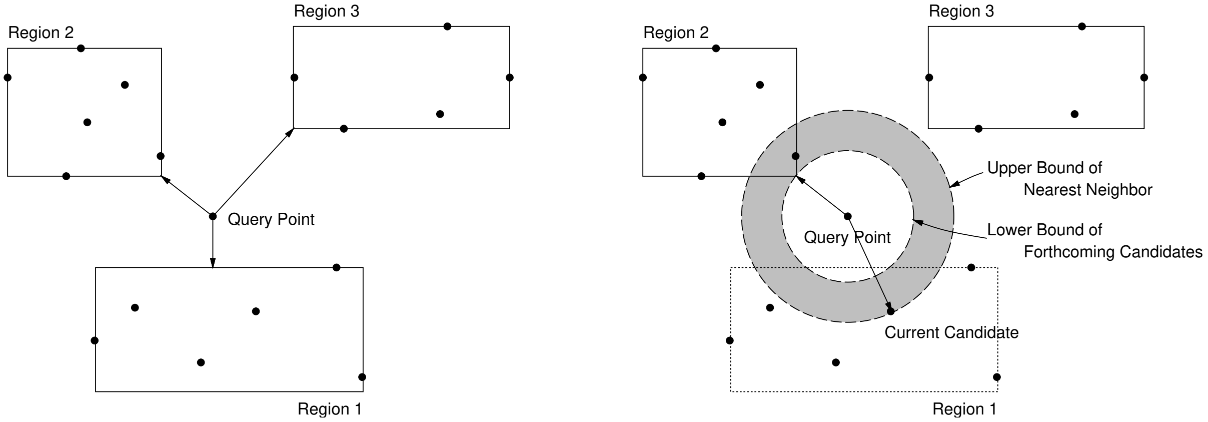
4 有意性感応型最近接点探索法

有意性の低い最近接点の影響を緩和するために、我々は、多次元インデックス構造のための新しい探索法を考案した。この探索法は、探索処理の実行中に、上述の判定法によって最近接点の有意性を判定し、最近接点の有意性が低いとわかると探索処理を中断し、近似解を返すというものである。我々はこの探索法のことを「有意性感応型最近接点探索 (significance-sensitive nearest neighbor search)」と呼んでいる。この探索法は、最近接点の有意性を評価できるだけでなく、後述のように、探索コストの低減も可能にする。

4.1 多次元インデックス構造の最近接点探索法

この論文が提案する有意性感応型最近接点探索法は、データ空間を領域の入れ子階層に分割する多次元インデックス構造（例えば、SS-tree[5], VAMSplit R-tree[6], X-tree[7], SR-tree[1, 2] など）を対象として設計したものである。図 6 に R-tree の階層構造を示す。木構造の各ノードは、データ空間の部分領域に対応する。非中間ノードは点データを格納し、中間ノードは子ノードに関する情報（領域に関する情報とポインタ）を格納する。階層的な構造によって、データ空間の一部の領域を探索するだけで最近接点探索が完了する。これにより、CPU 時間やディスク読み出しを減らすことが可能になる。そのため、このような多次元インデックス構造は、マルチメディア情報の類似検索を高速化する手段として広く使われている。

我々は、Hjaltason ら [8] の既存の最近接点探索法を拡張することで、新しい探索法を設計した（有意性感応型と区別するために、この探索法を基本型最近接点探索法と呼ぶことにする）。基本型最近接点探索法は、与えられた質問点に対する最近接点を、以下の手順で発見する。探索は、インデックス構造の根ノードから始まる。中間ノードを訪れる度に、質問点から子ノードまでの距離を計算し、子ノードはその距離の昇順に、優先順位付待ち行列に格納される。そして、待ち行列の先頭からノードをひとつ取りだし、そのノードに移動して探索を続ける。このように、ノードは、質問点からの距離の昇順に探索される。例えば、図 7 (a) では、Region 1、Region 2、Region 3 の順に探索される。一方、非中間ノードを訪れた時には、そのノードが格納している個々の点データについて、質問点からの距離を計算する。もし、これが非中間ノードの最初の訪問である場合には、そのノードで最も近い点が、最近接点の候補として選ばれる。そうでない場合には、そのノードで最も近い点と、それまでに得られている最近接点の候補とを比較し、より近い方を新しい候補として選ぶ。このとき、候補までの距離は、最近接点までの距離に対する上界を与える。なぜならば、最近接点は、候補に選ばれている点よりも遠い位置には存在し得ないからである。一方、優先順位付き待ち行列の先頭のノードまでの距離は、以後の探索で現れ得る新しい候補までの距離に対する下界を与える。なぜならば、待ち行列の先頭のノードは、まだ探索していないノードの中で最も近いノードだからである。例えば、図 7 (b) は、Region 1 が探索され、Region 1 の中で最も近い点が、最近接点の候補に選ばれた状態を示している。Region 2 と Region 3 は、優先順位付待ち行列の中に格納されたままである。この時点での候補が、最近接点までの距離の上界を与える一方、優先順位付待ち行列の先頭のノードまでの距離（すなわち、Region 2 までの距離）は、以後の探索で現れ得る候補までの距離の下界を与える。探索は、以後の探索で現れ得る候補までの距離の下界が、最近接点までの距離の上界を超えるまで（すなわち、現在の候補よりも近い距離にあるノードが、待ち行列中になくなるまで）続けられる。探索を終えた時点での候補が、真の最近接点である。以上の



(a) (b)

図 7: 多次元インデックス構造の最近接点探索法。

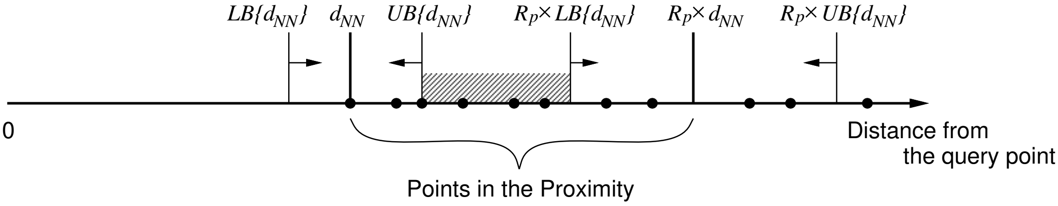


図 8: 近傍領域の下界と上界は、最近接点までの距離の下界と上界から求まる。

説明は、1 番めの最近接点の探索についてのものであるが、k 番めまでの最近接点の探索は、以上の手法を拡張することで容易に実現できる。

基本型最近接点探索は、真の最近接点を発見するが、与えられた精度のもとで、近似的に最近接点を探索する手法が提案されている。この最近接点探索は、近似的最近接点探索 (approximate nearest neighbor search)[9] と呼ばれている。近似的最近接点探索は、基本型最近接点探索の終了条件を変更することによって実現される。基本型最近接点探索が、現れ得る候補までの距離の下界が、最近接点までの距離の上界を超えた時点で終了するのに対して、近似的最近接点探索では、下界に対する上界の比が $(1+\epsilon)$ 以下になった時点で処理を終了する。ここに、ϵ は許容誤差の上限を定めるパラメータである。この時点で処理を終えると、質問点からの距離に関する誤差が ϵ 以下の範囲で、近似解を得ることができる。この近似的最近接点探索は、厳密解を必要としない場合には、探索処理の効率を改善することを可能にする。

4.2 有意性感応型最近接点探索法

有意性感応型最近接点探索法は、上記の基本型最近接点探索法を拡張したものである。最も重要な点は、探索中に、質問点の近傍領域に存在する点の数を数えることにある。近傍領域は、定義 1 のように、パラメータ R_p によって決定される。近傍領域における点の輻輳が検出されると、探索をその時点で終了し、近似解を返す。点の輻輳は、定義 1 のとおり、パラメータ N_c によって判定される。すなわち、近傍領域に存在する点の数が N_c を超えると、輻輳と見做されるのである。点の数を数えるために、この探索法では、最近接点までの距離 d_{NN} に対する下界と上界を利用する。基本型最近接点探索法で説明したとおり、待ち行列中の先頭のノードまでの距離と、最近接点の候補までの距離は、d_{NN} の下界と上界を与える (図 7 (b))。この下界と上界は、図 8 が示すように、近傍領域に対する下界と上界も与える。定義 1の近傍領域の定義より、近傍領域の範囲は、$R_p \times d_{NN}$ である。d_{NN} に対する下界と上界が、それぞれ $LB\{d_{NN}\}$、$UB\{d_{NN}\}$ であるとき、近傍領域の範囲に対する下界と上界は、$R_p \times LB\{d_{NN}\}$、$R_p \times UB\{d_{NN}\}$ として与えられる。d_{NN} は、探索が完了して初めて求まる値であるので、この探索法では、$UB\{d_{NN}\}$ から $R_p \times LB\{d_{NN}\}$ の範囲に存在する点の数を数える。図 8 のように、$d_{NN} \leq UB\{d_{NN}\}$ かつ $R_p \times LB\{d_{NN}\} \leq R_p \times d_{NN}$ であるから、近傍領域に存在する点の数は、$UB\{d_{NN}\}$ から $R_p \times LB\{d_{NN}\}$ までの範囲に存在する点の数に等しいかそれ以上である。したがって、$UB\{d_{NN}\}$ から $R_p \times LB\{d_{NN}\}$ までの範囲に、N_c 個以上の点が存在するかどうか調べることで、点の輻輳を、探索処理の実行中に検出することが可能になる。点の輻輳を検出した場合、その時点での候補を検索結果として返すことで、近似解を返すことができる。この探索法の利点は、点の輻輳によって最近接点の有意性が低いことが検出された場合に、真の最近接点の発見を放棄し、近似解を返すことにある。これにより、最近接点の有意性を識別で

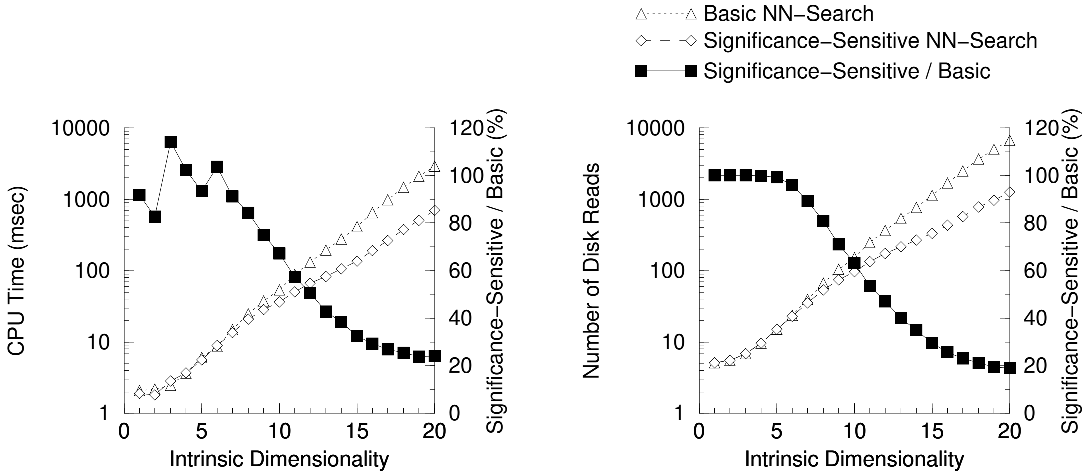


図 9: 合成データを用いた場合の探索コスト。

きるだけでなく、有意性の低い最近接点に対する処理を減らすことによって、探索コストの低減が可能になる。

5　評価実験

5.1　合成データによる評価

有意性感応型最近接点探索法の特性を評価するために、様々な埋め込み次元数を持つデータセットを合成した。埋め込み次元数 ν を持つデータセットは、n 次元の点 $(x_1, \ldots, x_n)$ を次式に従って生成することで合成した。

$$x_i = \begin{cases} U(0,1) & (1 \le i < \nu) \\ \dfrac{1}{\sqrt{n-\nu+1}}U(0,1) & (i = \nu) \\ x_\nu & (\nu < i \le n), \end{cases} \tag{9}$$

ここに、$U(0,1)$ は、0 から 1 の範囲の一様分布である。データ空間の次元数（すなわち、式 (9) の n ）を 20 として、埋め込み次元数が 1 から 20 のデータセットを合成した。ひとつのデータセットに含まれる点の数は 1,000,000 個である。

実験は、サンマイクロシステムズ社の Ultra 60 (CPU: UltraSPARC-II 360MHz, 主記憶: 512Mbytes, OS: Solaris 2.6) の上で行った。プログラムは C++ によって書かれており、インデックス構造としては、優れた静的構築アルゴリズムを持つ VAMSplit R-tree[6] を用いた。1 番めの最近接点を見つける処理を、無作為に選んだ 1,000 個の質問点に対して実行し、平均値を測定結果とした。質問点は、無作為にデータセットの中から選んでいる。パラメータ R_p と N_c は、3.2 節で述べた設定法に従い 1.84471 と 48 に設定した。

図 9 と 図 10 に実験結果を示す。どちらの図についても、横軸は、データセットの埋め込み次元数である。図 9 は、CPU 時間とディスク読み込み回数を示している。埋め込み次元数が低いときには、基本型最近接点探索法と有意性感応型最近接点探索法の差はほとんど見られない。しかし、埋め込み次元数が高くなると、CPU 時間、ディスク読み込み回数とも著しく減少している。埋め込み次元数が 20 のときには、CPU 時間が 76%、ディスク読み込み回数が 81% 減少している。この結果は、最近接点探索の処理コストが、有意性感応型最近接点探索によって低減できることを示している。図 10 は、最近接点の棄却率（すなわち、最近接点の有意性が低いと判定された割合）を示している。測定された棄却率は、理論曲線と非常に近いことがわかる。この結果は、3.2節で述べたパラメータの設定法の妥当性を示している。

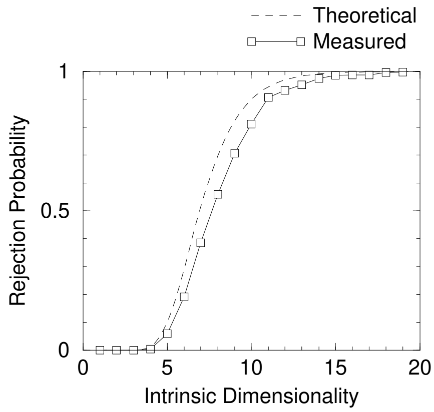


図 10: 合成データを用いた場合の棄却率。

5.2　実データによる評価

有意性感応型最近接点探索法を静止画像の類似検索に適用し、その有効性を検証した。NASA の写真および映像のアーカイブから 40,745 枚の画像を集め、カラーヒストグラムを特徴ベクトルとした。色空間としては、マンセル表色系を使い、9 色の部分空間（黒、灰、白、ならびに、色相で分割した 6 色）に分割した。画像の構図を反映させるために、個々の画像を 4 分割し、それぞれの領域についてカ

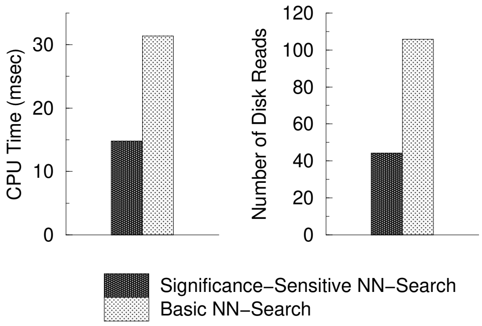


図 11: 実データを用いた場合の探索コスト。

図 12: 検索結果の例。

ラーヒストグラムを求め、それらを結合して 36 次元の特徴ベクトルを算出した。そしてさらに、主成分分析によって次元を 20 にまで減らし、特徴空間とした。特徴ベクトル間の類似度は、ユークリッド距離によって評価した。全ての画像の特徴ベクトルを質問点として、10 番めまでの最近接点を探索する処理を実行し、平均値を測定結果とした。従って、見つかった最近接点の中のひとつは、質問点そのものである。

図 11 に、基本型最近接点探索と有意性感応型最近接点探索の処理コストを示す。CPU 時間、ディスク読み込み回数とも、有意性感応型最近接点探索の方が減少している。CPU 時間で 48%、ディスク読み込み回数で 56% の減少である。図 12 は、検索結果の一例である。個々の画像の下の数値は、質問点からの距離である。有意性感応型最近接点探索は、No. 6 の画像の有意性が低いことを検出した。そして、No. 1 ∼ 5 については厳密解を、No. 6 ∼ 10 については近似解を返している。この検索結果から、No. 2 から No. 5 の画像の方が、No. 6 から No. 10 の画像よりも、有意性が高いことを知ることができる。この例からわかるとおり、有意性感応型最近接点探索法は、ユーザが最近接点を有意性の高いものと低いものとに分類することを可能にする。この特長は、対話的な検索システムにおいて特に有効であると考えられる。

6 むすび

この論文では、マルチメディア情報の効率的な検索法として、有意性感応型最近接点探索法を提案した。高次元空間では、有意性の低い最近接点が起こりやすく、最近接点探索の処理効率を下げる原因になる。この探索法は、最近接点の有意性を評価できるだけでなく、従来の最近接点探索法に比べて、探索コストの低減も可能にする。これらの利点は、対話的な検索システムにおいて、特に有効であると考えられる。今後は、他のデータや他の特徴量に対しても提案手法を適用し、その特性ならびに有効性を検証していく考えである。

謝辞

本研究の一部は、文部省創成的基礎研究費（09NP1401）、および、文部省科学研究費補助金（奨励 12780251）の補助を受けた。

参考文献

[1] N. Katayama and S. Satoh, “The SR-tree: An Index Structure for High-Dimensional Nearest Neighbor Queries,” Proc. of the 1997 ACM SIGMOD, Tucson, USA (May 1997) pp. 369–380.

[2] 片山紀生，佐藤真一，“SR-Tree: 高次元点データに対する最近接検索のためのインデックス構造の提案，” 信学論（D-I），vol.J80-D-I，no.8，pp.703–717，Aug. 1997．

[3] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When Is “Nearest Neighbor” Meaningful?,” Proc. of the 7th Int. Conf. on Database Theory, Jerusalem, Israel, pp.217–235, Jan. 1999.

[4] K. Fukunaga, “Introduction to Statistical Pattern Recognition (2nd ed.),” Academic Press, 1990.

[5] D. A. White and R. Jain, “Similarity Indexing with the SS-tree,” Proc. of the 12th Int. Conf. on Data Engineering, New Orleans, USA, pp.516–523, Feb. 1996.

[6] D. A. White and R. Jain, “Similarity Indexing: Algorithms and Performance,” Proc. SPIE Vol.2670, San Diego, USA, pp.62–73, Jan. 1996.

[7] S. Berchtold, D. A. Keim, and H.-P. Kriegel, “The X-tree: An Index Structure for High-Dimensional Data,” Proc. of the 22nd VLDB Conf., Bombay, India, pp.28–39, Sep. 1996.

[8] G. Hjaltason and H. Samet, “Ranking in Spatial Databases,” 4th Int. Symp. on Spatial Databases, SSD’95, Portland, USA, pp.83–95, Aug. 1995.

[9] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, “An Optimal Algorithm for Approximate Nearest Neighbor Searching,” Proc. of the 5th Ann. ACM-SIAM Symposium on Discrete Algorithms, pp.573–582, 1994.