確率的空間問合せ処理の効率化

飯島 裕一 †　　石川 佳治 ‡†
†名古屋大学大学院情報科学研究科　　‡名古屋大学情報基盤センター

1　はじめに

近年，位置情報サービスの普及に伴い，位置の曖昧さを考慮した空間問合せ処理技術の必要性が高まっている．これは，現実世界のオブジェクトの位置は曖昧にしか得られない場合が多いことに起因しており，例えば移動ロボット分野では，しばしばセンサ信号や移動履歴に基づく自己位置推定が行われるが，測定誤差や制御ノイズなどのために誤差を伴った推定となる．

本研究では，位置が曖昧なオブジェクトが，自身から最も近いオブジェクトを求めるために最近傍問合せを行うという状況を対象とする．具体的には，問合せを行う，問合せオブジェクトの位置が**正規分布**により確率的に表現され，問合せの対象となる，データオブジェクトが確定的な位置を持つ点データである場合を扱う．この場合の最近傍オブジェクトは確率的に決まるため，その確率により結果が定まるような問合せを定義する必要がある．そこで本研究では，**確率的最近傍問合せ**を定義し，その効率的な処理手法を提案する．

2　確率的最近傍問合せの定義

d次元空間中で，問合せオブジェクトqの位置がd次元ベクトルの座標値$\boldsymbol{x}$を持つ確率が，d次元正規分布の確率密度関数により，下式で表現されるとする．

$$p_q(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{q})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{q})\right] \quad (1)$$

このとき，qの最近傍オブジェクトとなる確率がθ $(0<\theta<1)$以上であるオブジェクトの集合を返す問合せを確率的最近傍問合せ$PNNQ(q,\theta)$と定義する．データオブジェクトの集合を$\mathcal{O}$として，$o\in\mathcal{O}$がqの最近傍オブジェクトとなる確率$\Pr_{NN}(q,o)$は下式で表される．

$$\Pr_{NN}(q,o)=\Pr(\forall o'\in\mathcal{O}, o'\neq o, \|\boldsymbol{x}-\boldsymbol{o}\|^2\leq\|\boldsymbol{x}-\boldsymbol{o}'\|^2) \quad (2)$$

これを用いて$PNNQ(q,\theta)$の式は以下のとおり表せる．

$$PNNQ(q,\theta)=\{n \mid n\in\mathcal{O}, \Pr_{NN}(q,n)\geq\theta\} \quad (3)$$

問合せの入力として与えられるのは，qの情報（$p_q(\boldsymbol{x})$の共分散行列$\boldsymbol{\Sigma}$と分布の平均$\boldsymbol{q}$）と確率の閾値θである．

3　提案手法

本手法では**ボロノイ図**を用いる．ボロノイ図は，図1に示すように，空間中の複数の点のうち，どの点に一番近いかによって空間を分割した図である．各点の勢力範囲は**ボロノイ領域**と呼ばれ，ボロノイ図の定義から，oのボロノイ領域V_o内にqが位置するときにoは

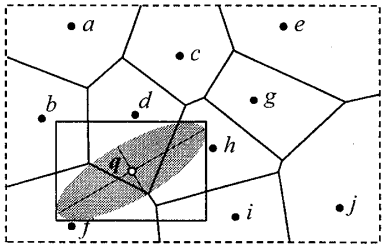

図1　ボロノイ図とθ-領域の包囲矩形

qの最近傍オブジェクトとなる．つまり，$\Pr_{NN}(q,o)$はqがV_o内に位置する確率に等しく，下式で計算できる．

$$\Pr_{NN}(q,o)=\int_{\boldsymbol{x}\in V_o} p_q(\boldsymbol{x})d\boldsymbol{x} \quad (4)$$

ただし，この計算にはコストの高い数値積分が必要となるため，すべてのデータオブジェクトに対して$\Pr_{NN}(\cdot)$を計算することは現実的ではない．そこで本手法では，明らかに$\Pr_{NN}(\cdot)$がθに満たないといえるオブジェクトを除去（フィルタリング）し，残ったオブジェクト（以降，候補オブジェクト）に対してのみ$\Pr_{NN}(\cdot)$を数値積分により計算してθ以上のものを出力する．このアプローチに基づく問合せ戦略を2種類提案する．

3.1　問合せ戦略1

本戦略では，**θ-領域**[1]を用いてフィルタリングを行う．θ-領域は，その領域での$p_q(\boldsymbol{x})$の積分値が$1-2\theta$になるような楕円体領域$(\boldsymbol{x}-\boldsymbol{q})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{q})\leq r_\theta^2$である．この関係を式で表すと以下のようになる．

$$\int_{(\boldsymbol{x}-\boldsymbol{q})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{q})\leq r_\theta^2} p_q(\boldsymbol{x})d\boldsymbol{x}=1-2\theta \quad (5)$$

図1を用いて本戦略のアイデアを説明する．図の陰影部分がθ-領域であるが，これを直接フィルタリングに用いることは難しいため，座標軸に平行な包囲矩形を考える．ここで，ボロノイ領域が包囲矩形と重なりを持たないオブジェクト，すなわちa,c,e,g,jを解の候補から除くことができる．その理由を以下に述べる．まず，θ-領域の定義から，包囲矩形の外側の領域での$p_q(\boldsymbol{x})$の積分値は$1-(1-2\theta)=2\theta$未満である．また，$p_q(\boldsymbol{x})$の分布は平均$\boldsymbol{q}$について点対称であるため，ボロノイ領域V_oと$\boldsymbol{q}$について対称な領域V_o'での$p_q(\boldsymbol{x})$の積分値はV_oでの積分値に等しい．これらの事実により，包囲矩形と重なりを持たないボロノイ領域での$p_q(\boldsymbol{x})$の積分値は2倍しても2θ未満ということになる．したがって，ボロノイ領域が包囲矩形と重なりを持たないオブジェクトは$\Pr_{NN}(\cdot)$がθ以上になることはない．

問合せ時に与えられるθに応じて，θ-領域の包囲矩形を動的に導出する方法については[1]を参照されたい．

Improving Efficiency of Probabilistic Spatial Query Processing
Yuichi Iijima †　Yoshiharu Ishikawa ††
†Graduate School of Information Science, Nagoya University
‡Information Technology Center, Nagoya University

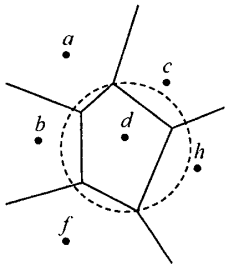

図2 最小包含球

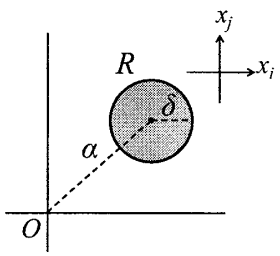

図3 球領域 R

α	δ	$\pi(\alpha,\delta)$
0.0	0.1	⋯
0.0	0.2	⋯
⋮	⋮	⋮
1.0	0.1	⋯
1.0	0.2	⋯
⋮	⋮	⋮

図4 対応表

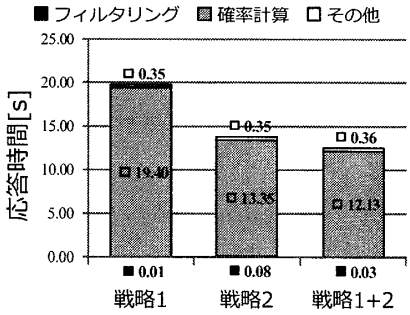

図5 応答時間とその内訳

3.2 問合せ戦略2

本戦略では，各データオブジェクトに対して，ボロノイ領域の代わりにその最小包含球の領域での $p_q(\boldsymbol{x})$ の積分値を求める．例としてボロノイ領域 V_d の最小包含球 SES_d を図2に示す．最小包含球の領域での積分値は $\Pr_{NN}(\cdot)$ より大きくなるため，この値が θ 以下のオブジェクトは棄却できることになる．

任意の球領域での積分値は事前に対応表を作成しておくことで簡単に求められる．まず，$\boldsymbol{\Sigma}$ が単位行列であるという単純な場合について説明する．この場合の $p_q(\boldsymbol{x})$ は，標準正規分布の確率密度関数 $p_{\mathrm{norm}}(\boldsymbol{x})$ を $\boldsymbol{q}$ が中心となるように平行移動したものに等しくなる．$p_{\mathrm{norm}}(\boldsymbol{x})$ の等確率面は球形のため，原点からの距離と半径がともに等しい任意の球領域での積分値は一定である．そこで，原点から距離 α の点を中心とする半径 δ の d 次元の球領域 R（図3の陰影部分）での，$p_{\mathrm{norm}}(\boldsymbol{x})$ の積分値を以下のように表す．

$$\pi(\alpha,\delta) = \int_{\boldsymbol{x}\in R} p_{\mathrm{norm}}(\boldsymbol{x})d\boldsymbol{x} \qquad (6)$$

異なる α と δ の値の組合せに対して，数値積分により $\pi(\alpha,\delta)$ を計算することで，図4のような表を作成する．$\int_{\boldsymbol{x}\in SES_o} p_q(\boldsymbol{x})d\boldsymbol{x}$ の値は，$\boldsymbol{q}$ から SES_o の中心までの距離を α_o，SES_o の半径を δ_o としたときの $\pi(\alpha_o,\delta_o)$ に等しいため，(α_o,δ_o) のエントリを表から検索すれば得られる．得られた値が θ 以下の場合，o を棄却できる．

$\boldsymbol{\Sigma}$ が任意の場合，$p_q(\boldsymbol{x})$ の等確率面は楕円体の形状をとるため，単純に (α,δ) の表から任意の球領域での積分値を得ることはできない．そこで，$p_q(\boldsymbol{x})$ の代わりに，等確率面が球形であり，任意の $\boldsymbol{x}$ に対して $p_q(\boldsymbol{x}) \le p_q^{\top}(\boldsymbol{x})$ が成立する上限の関数 $p_q^{\top}(\boldsymbol{x})$[1] の最小包含球領域での積分値を求める．この値は同じ領域での $p_q(\boldsymbol{x})$ の積分値より大きくなるため，値が θ 以下のオブジェクトは棄却できる．また，この値は $\boldsymbol{\Sigma}$ が単位行列の場合に作成した図4の表から求められるが，詳細は割愛する．

4 評価実験

4.1 実験方法

2つの戦略にそれらのハイブリッド戦略を加えた3つの戦略の性能を評価した．ハイブリッド戦略は，始めに戦略1のフィルタリングを行い，残ったオブジェクトに対して戦略2のフィルタリングを行う戦略である．

表1 候補オブジェクト数及び解オブジェクト数

戦略1	戦略2	戦略1+2	解
101.5	62.3	55.0	5.4

データオブジェクトとして，米国ロングビーチの道路データから作成された50,501個の2次元点を用いて，$\boldsymbol{q}$ の異なる100回の問合せの平均応答時間を調べた．

$p_q(\boldsymbol{x})$ の等確率線の形状と大きさは $\boldsymbol{\Sigma}$ によって決定される．今回は標準の形状を，長軸と短軸の比が3:1で傾き30°の楕円とした．また，大きさは位置の曖昧さの程度に対応しており，適正な評価ができるように標準の設定を決定した．θ は0.03を標準とした．

4.2 実験結果

標準の設定における結果を図5及び表1に示す．各戦略の処理時間の大部分は $\Pr_{NN}(\cdot)$ の計算に費やされており，フィルタリングにより $\Pr_{NN}(\cdot)$ の計算を行うオブジェクトを減らすアプローチの正しさが確認された．

位置の曖昧さの程度，θ，$p_q(\boldsymbol{x})$ の等確率線の形状をそれぞれ標準の設定から変動させた場合についても実験を行ったところ，基本的には戦略2の方が性能が良いが，戦略1に対する優位性の程度は設定に依存することがわかった．特に，曖昧さが大きい，θ が大きい，$p_q(\boldsymbol{x})$ の等確率線の形状が球形に近い，という状況では戦略2の優位性が高かった．ただし，実用上は最も良い性能を示したハイブリッド戦略を用いるべきである．

5 まとめ

本研究では，位置の曖昧さが正規分布で表現されたオブジェクトが確定的な位置を持つオブジェクトを対象に行う最近傍問合せの処理手法を提案した．明らかに解でないといえるオブジェクトを求めるというアプローチに基づいて，2種類の戦略を提案し，評価を行った．

謝辞

本研究の一部は，文部科学省科学研究費（19300027, 21013023）の助成による．

参考文献

[1] Y. Ishikawa, Y. Iijima, and J. X. Yu. Spatial range querying for Gaussian-based imprecise query objects. In *Proc. ICDE*, pp. 676–687, 2009.