社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－146 (40)
2008／9／22

曖昧な位置に基づく最近傍問合せ処理手法

飯島 裕一† 石川 佳治††

† 名古屋大学大学院情報科学研究科 〒464–8601 名古屋市千種区不老町
†† 名古屋大学情報連携基盤センター 〒464–8601 名古屋市千種区不老町
E-mail: †iijima@db.itc.nagoya-u.ac.jp, ††ishikawa@itc.nagoya-u.ac.jp

あらまし センサ環境や移動ロボットなどの分野では，センサの測定誤差やオブジェクト自身の移動などのために，オブジェクトの位置が曖昧なものとなる状況がしばしば生じる．本稿では，問合せオブジェクトの位置が正規分布の確率密度関数によって曖昧な位置で表現されている状況における最近傍問合せの処理手法について述べる．
キーワード 空間データベース，最近傍問合せ，曖昧な位置，正規分布，センサデータベース

Nearest-Neighbor Query Processing Methods Based on Imprecise Locations

Yuichi IIJIMA† and Yoshiharu ISHIKAWA††

† Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464–8601 Japan
†† Information Technology Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464–8601 Japan
E-mail: †iijima@db.itc.nagoya-u.ac.jp, ††ishikawa@itc.nagoya-u.ac.jp

Abstract In sensor environments and moving robot applications, the position of an object is often imprecise due to the measurement error in the sensor and the object's own movement. In this paper, we present a query processing technique for nearest-neighbor queries in which the position of the query object is represented by a Gaussian distribution.
Key words spatial databases, nearest-neighbor queries, imprecise locations, Gaussian distributions, sensor databases

1. はじめに

近年，センサ環境や移動ロボットなどの分野において，曖昧な位置情報に基づくデータベース問合せ処理技術の必要性が高まってきている．多くのセンサが各々の周辺の環境情報を収集するセンサ環境では，個々のセンサが「どこの情報を計測しているのか」という情報はなくてはならないものである．各センサの位置を把握する手段として GPS を利用する方法が考えられるが，電波状況によっては期待通りの測位精度を得られないという問題がある．加えて，GPS による位置取得は多くの電力を消費するため，各センサが電池で駆動しているような場合には極力避けたいという要求もある．また，現実環境中を動き回る移動ロボットにとって，自身の位置の推定は円滑なサービス提供を行う上で欠かせないものであるが，センサの分解能やモータの制御ノイズなどのために，正確な位置の推定は容易ではなく，誤差を伴った推定となる．以上のように，センサ環境における各センサの位置や移動ロボットの位置は曖昧であるため，そのことを踏まえた問合せ処理手法が必要とされており，また，そのような曖昧なデータを対象とした問合せに関する研究が最近盛んになってきている [1].

このような背景から，本研究では，その位置が曖昧な位置で表現されているオブジェクトが，自らの位置に最も近い距離にあるオブジェクトを検索するために最近傍問合せを行うという状況を対象とし，その問合せ処理手法を提案する．具体的には，問合せオブジェクトの位置が**正規分布**で表現され，問合せ対象オブジェクトが確定的な位置で表される**点データ**である状況を想定している．対象とする問合せとしては，ユークリッド距離に基づく通常の最近傍問合せを拡張した**確率的最近傍問合せ**を考えることにするが，詳しくは 2 節で述べる．本研究グループでは，すでに，本研究が対象とする状況と同様の状況における範囲問合せについて，その処理手法を [2] で提案している．本研究が対象とするのは最近傍問合せであり，[2] とは対象とする問合せが異なるが，後に 3 節で説明する上限の関数を考えるアプローチなど，共通している部分も多い．

本稿の構成は以下の通りである．まず，2 節で本研究が対象とする問合せである確率的最近傍問合せについて述べる．次に，3 節でその処理手法を提案する．最後に，4 節でまとめを行うとともに今後の課題を挙げる．

2. 確率的最近傍問合せ

2.1 最近傍問合せ

本研究では，対象とする問合せとして，通常の最近傍問合せを拡張した確率的最近傍問合せを考える．そこで，まずはユークリッド距離に基づく通常の最近傍問合せについて説明する．最近傍問合せとは，指定された点に最も近い位置にあるオブジェクトを求める問合せのことである．各オブジェクトが固定的な位置で表されているような状況においては，最近傍問合せを考える上で何の問題も起こらない．しかしながら，本研究が対象とする，問合せオブジェクトの位置が曖昧な位置で表現されている状況というのは，通常の最近傍問合せでは対応することができない状況である．そのため，本研究では，通常の最近傍問合せを拡張した確率的最近傍問合せを定義することにする．このことについて，1 次元の場合を例に次節で説明する．

2.2 1 次元の確率的最近傍問合せ

例 1 1 次元直線上に問合せオブジェクト q と問合せ対象オブジェクト a, b, c, d, e が存在している状況を考える．a, b, c, d, e は図 1 に示すような位置に固定されており，その位置に曖昧性はないとする．一方，q の位置は不定であり，その x 座標 x_q の確率密度関数 $p_q(x)$ が原点を中心とする分散 1 の正規分布 $\mathcal{N}(\mu, \sigma^2) = (0, 1)$ で表されているとする．図中の曲線はその確率密度曲線を示している．

このとき，q が自身の位置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発行したとする．仮に，q の位置が曖昧ではなく，$x = 0$ に固定されているとすれば，問合せ結果は $\{c\}$ ということになるが，この例のように，q の位置が確率的に指定されている状況では，問合せ結果を一意に決めることはできない．なぜなら，q は直線上のどこにでもある可能性があるからである．例えば，q の位置が $x = -1$ であるとすれば問合せ結果は $\{b\}$ となるが，$x = 1$ であるとすれば問合せ結果は $\{d\}$ となる．最近傍問合せの問合せ結果は問合せオブジェクトと問合せ対象オブジェクトの位置関係によって決まるため，問合せオブジェクトの位置が確率的である場合，問合せ結果も確率的になるのである．

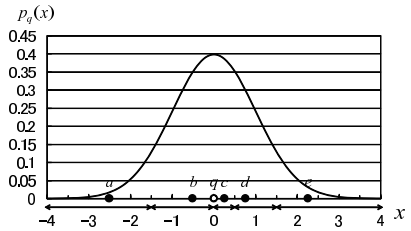

図 1 1 次元の確率的最近傍問合せ

そこで，このような状況に対応するため，通常の最近傍問合せの概念を拡張した**確率的最近傍問合せ**（**probabilistic nearest neighbor query, PNNQ**）を以下のように定義する．

定義 1 q を，その位置が $\mathcal{N}(\mu, \sigma^2)$ に従うオブジェクトとする．すなわち，q の位置が x である確率が，確率密度関数

$$p_q(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{1}$$

で表されるとする．つまり，$p_q \sim \mathcal{N}(\mu, \sigma^2)$ である．このような q が与えられたとき，q とのユークリッド距離がすべてのオブジェクトのうちで最も小さくなる確率が θ 以上であるようなオブジェクトの集合を $PNNQ(q, \theta)$ で表す．ただし，θ は $0 < \theta < 1$ を満たすものとする．$PNNQ(q, \theta)$ を式で表現すると以下のようになる．

$$PNNQ(q, \theta) = \{n \mid n, \forall o \in \mathcal{O},\ o \neq n, \Pr(dist(x_q, x_n) \leqq dist(x_q, x_o)) \geqq \theta\} \tag{2}$$

ただし，$\mathcal{O}$ は問合せ対象オブジェクトの集合であり，$dist(x_q, x_o)$ は問合せオブジェクト q の x 座標 x_q と問合せ対象オブジェクト o の x 座標 x_o のユークリッド距離である．

通常の最近傍問合せとの違いは，確率の閾値 θ を導入しているところである．本研究の対象とする状況では，各問合せ対象オブジェクトが最近傍問合せを満たすかどうかは確率的に決まるため，その確率値がユーザの与える閾値 θ 以上であるかどうかによって，そのオブジェクトが問合せ結果に含まれるか否かを決定する．なお，θ を $0 < \theta < 1$ としたのは，問合せオブジェクトの位置が正規分布の確率密度関数で表現されているということに関係がある．正規分布は分布が無限遠まで拡がっているため，どの問合せオブジェクトについても，そのオブジェクトが問合せオブジェクトの最近傍オブジェクトとなる確率は 0 より大きいことになる．よって，$\theta = 0$ とすると，すべての問合せ対象オブジェクトが問合せ結果に含まれることになる．また，$\theta = 1$ とすると，$\theta = 0$ の場合とは逆に，どの問合せ対象オブジェクトも問合せ結果に含まれないことになる．つまり，$\theta = 0$ または $\theta = 1$ としても意味のある問合せにはならないため，θ を $0 < \theta < 1$ としている．

定義 1 により，例えば，問合せオブジェクト q から最も近くに位置している確率が 20%以上であるオブジェクトの集合を求める問合せは，$PNNQ(q, 0.2)$ を求める確率的最近傍問合せと言い換えることができる．ここで，例を用いて確率的最近傍問合せの処理について説明する．

例 2 例 1 において，q が $PNNQ(q, 0.2)$ を求める確率的最近傍問合せを発行したとする．この問合せを処理するためには，各問合せ対象オブジェクトについて，そのオブジェクトが q から最も近いオブジェクトとなる確率を求め，その値が 0.2 以上かどうかを評価すればよい．例えば，問合せ対象オブジェクト a について，a が q から最も近い位置にあるオブジェクトとなるのは，

q の位置が $x_q \leqq -1.5$ の範囲にある場合である．この範囲内に q が位置する確率は x_q の確率密度関数をこの範囲で積分することで得られ，計算すると，$\int_{-\infty}^{-1.5} p_q(x)dx = 0.067$ となる．よって，a が q の最近傍オブジェクトとなる確率は $\Pr(a) = 0.067$ である．同様に，b が q の最近傍オブジェクトとなるのは $-1.5 \leqq x_q \leqq 0$ の場合であり，その確率は $\Pr(b) = \int_{-1.5}^{0} p_q(x)dx = 0.433$ である．c, d, e についても q の最近傍オブジェクトとなる確率をそれぞれ計算すると，$\Pr(c) = \int_{0}^{0.5} p_q(x)dx = 0.191$，$\Pr(d) = \int_{0.5}^{1.5} p_q(x)dx = 0.242$，$\Pr(e) = \int_{1.5}^{\infty} p_q(x)dx = 0.067$ となる．以上の結果から，$\theta = 0.2$ という閾値以上となるのは b, d のみである．よって，問合せ結果は $\{b, d\}$ と求まる．

上記の例の場合では，正規分布の平均，すなわち，q の平均の位置 $x_q = 0$ に最も近い位置にあるオブジェクトである c が問合せ結果に含まれず，c よりも q の平均の位置から遠い b, d が問合せ結果に含まれるという結果となった．これは，それぞれのオブジェクトの「q の最近傍オブジェクトとなるような範囲」（以下,「勢力範囲」と呼ぶことにする）が，b は $-1.5 \leqq x_q \leqq 0$ で大きさ 1.5，d は $0.5 \leqq x_q \leqq 1.5$ で大きさ 1 であるのに対して，c は $0 \leqq x_q \leqq 0.5$ で大きさ 0.5 と小さいことが影響している．それぞれのオブジェクトの勢力範囲を図 1 の x 軸の下に矢印で示した．a, e は b, d よりも勢力範囲の大きさが大きいが，上記の例の問合せを満たさない．これは，b, d に比べて a, e が q の平均の位置 $x = 0$ から遠い位置にあることが影響している．結局のところ，確率的最近傍問合せにおいて，あるオブジェクトが問合せ結果に含まれるかどうかを決める上での重要な要素は，そのオブジェクトの勢力範囲の大きさと，そのオブジェクトの問合せオブジェクトの平均の位置からの距離である．

2.3 多次元の確率的最近傍問合せ

前節で定義した確率的最近傍問合せは 1 次元の場合についてのものであった．本節では，これを d 次元（$d \geqq 2$）の場合に対応できるように拡張する．d 次元の場合について，確率的最近傍問合せを定義しなおすと以下のようになる．

定義 2 d 次元空間において，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 次元正規分布により，

$$p_q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{q}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{q})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{q})\right] \quad (3)$$

で表されるとする．ただし，$\boldsymbol{\Sigma}$ は $d \times d$ の共分散行列であり，$|\boldsymbol{\Sigma}|$ は $\boldsymbol{\Sigma}$ の行列式を表す．$\boldsymbol{q}$ は確率分布の平均の位置に対応している．このような q が与えられたとき，q とのユークリッド距離がすべてのオブジェクトのうちで最も小さくなる確率が θ 以上であるようなオブジェクトの集合を $PNNQ(q, \theta)$ で表す．ただし，θ は $0 < \theta < 1$ を満たすものとする．$PNNQ(q, \theta)$ を式で表現すると以下のようになる．

$$PNNQ(q, \theta) = \{n \mid n, \forall o \in \mathcal{O},\ o \neq n, \Pr(\|\boldsymbol{x}-\boldsymbol{n}\|^2 \leqq \|\boldsymbol{x}-\boldsymbol{o}\|^2) \geqq \theta\} \quad (4)$$

ただし，$\mathcal{O}$ は問合せ対象オブジェクトの集合であり，$\|\cdot\|$ はベクトルの長さを表す．$\|\boldsymbol{x}-\boldsymbol{o}\|^2$ は，問合せオブジェクト q の位置 $\boldsymbol{x}$ と問合せ対象オブジェクト o の位置 $\boldsymbol{o}$ のユークリッド距離の 2 乗を表している．

以上のように定義した確率的最近傍問合せの処理手法を次節で提案する．

3. 問合せ処理手法

3.1 基本的なアイデア

d 次元の確率的最近傍問合せを処理するために，まずは，2.2 節で示した 1 次元の場合と同様に，各問合せ対象オブジェクトに対して，そのオブジェクトが問合せオブジェクトの最近傍オブジェクトとなるような領域を求める必要がある．つまり，各オブジェクトの「勢力範囲」を求める必要があるということである．本手法では，各オブジェクトの勢力範囲を求めるために，ボロノイ図を利用する．**ボロノイ図** [3] とは，空間上にいくつかの点が与えられた際，どの点に一番近いかによって空間を分割した図のことである．近隣の点同士を結んだ直線に対する垂直二等分線をつなげることによって作図することができる．図 2 に 2 次元平面上のオブジェクト a, b, c, d, e に対するボロノイ図を示す．各オブジェクトの勢力範囲を**ボロノイ領域**という．例えば，図 2 の灰色部分で示した領域が c のボロノイ領域である．これまで「勢力範囲」と呼んでいた領域は，まさにこのボロノイ領域のことである．c のボロノイ領域に問合せオブジェクト q が位置している場合に，c は q の最近傍オブジェクトとなる．よって，c が q の最近傍オブジェクトとなる確率は，c のボロノイ領域で q の位置についての確率密度関数を積分することで求められる．この確率が閾値以上であれば，c は問合せ結果に含まれるということになる．

本研究では問合せオブジェクトの位置が正規分布で表現されている状況を対象としているため，ある問合せ対象オブジェクトが問合せオブジェクトの最近傍オブジェクトとなる確率を求めるためには，そのオブジェクトのボロノイ領域で正規分布の確率密度関数，すなわち，式 (3) を積分した値を計算する必要がある．しかしながら，正規分布の確率密度関数の積分値を解析的に求めることはできないため，コストの高い数値積分が必要となる．また，各ボロノイ領域はそれぞれ凸多角形という複雑な形状をとるため，計算コストはさらに高まる．よって，すべての問合せ対象オブジェクトに対して，そのオブジェクトが問合せオブジェクトの最近傍オブジェクトとなる確率を求めていると，コストが非常に高くなるという問題がある．そこで，本手法では，積分計算を行って具体的な確率値を求めるまでもなく，明らかに確率値が閾値 θ より小さいといえるオブジェクトをあらかじめ排除することによって，コストの削減を図ることにする．本手法のアイデアを，2 次元の場合を例に以下に示す．

まず，それぞれのボロノイ領域に対して，そのボロノイ領域を囲む最小の円（**最小包囲円** [3]）を求める．例として，オブジェクト c のボロノイ領域の最小包囲円を図 3 に示した．次に，求めた円の領域で正規分布の確率密度関数を積分する．すると，

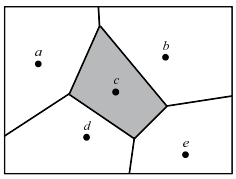

図 2　ボロノイ図

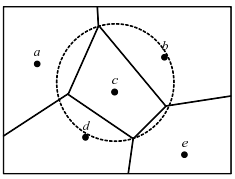

図 3　c のボロノイ領域の最小包囲円

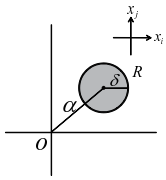

図 4　超　球　R

表 1　(α, δ) に対する $p_{\mathrm{norm}}(\boldsymbol{x})$ の積分値の対応表

α	δ	積分値
1.0	0.1	...
1.0	0.2	...

この積分値は，ボロノイ領域で正規分布の確率密度関数を積分した値の近似値とみなすことができ，なおかつ，ボロノイ領域で積分した値よりも大きな値となる．さらに，詳しくは後述するが，正規分布の確率密度関数を円の領域で積分した値は，事前に表を作成しておくことによって簡単に求めることができる．よって，最小包囲円の領域での積分値が閾値未満のオブジェクトについては，コストの高いボロノイ領域での積分を行うことなく，問合せを満たす可能性がないオブジェクトとして棄却することができるので，コストの削減につながる．

このアイデアは d 次元の場合について適用可能な，一般的なものである．3 次元の場合には，最小包囲円ではなく最小包囲球を用いればよい．同様に，4 次元以上の場合には最小包囲超球を用いればよい．

本手法において重要な役割を果たすのが，円（球，超球）の領域で正規分布の確率密度関数を積分した値を素早く導出するために使用する表である．この表について，以降の節で詳しく説明する．まず，次の 3.2 節で，式 (3) の共分散行列 $\boldsymbol{\Sigma}$ が単位行列の定数倍であるという，特殊な状況の場合について述べ，続く 3.3 節で一般の場合について述べる．

3.2　$\boldsymbol{\Sigma} = \sigma\mathbf{I}$ の場合

本節では，式 (3) の共分散行列が単位行列の定数倍である場合，すなわち，$\boldsymbol{\Sigma} = \sigma\mathbf{I}$ の場合について考える．σ は 0 より大きい定数である．このとき，$|\boldsymbol{\Sigma}| = \sigma^d$ となる．式 (3) において，$|\boldsymbol{\Sigma}| = \sigma^d, \boldsymbol{x} = (x_1, x_2, \ldots, x_d), \boldsymbol{q} = (q_1, q_2, \ldots, q_d)$ とおくと，

$$p_q(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sigma^{d/2}} \exp\left[-\frac{1}{2\sigma}\|\boldsymbol{x} - \boldsymbol{q}\|^2\right] \qquad (5)$$

となり，等確率面は超球の形状となる．$\|\boldsymbol{x} - \boldsymbol{q}\|^2 = \sum_{i=1}^{d}(x_i - q_i)^2$ はユークリッドノルムの 2 乗である．

各オブジェクトのボロノイ領域はその大きさも位置もさまざまである．よって，そのボロノイ領域を囲う最小包囲円（球，超球）についても，その半径や中心点の座標はオブジェクトごとに異なる．異なる半径，異なる中心点の座標を持つさまざまな円に対して，その円の領域で $p_q(\boldsymbol{x})$ を積分した値を素早く導出できるように，表を以下のようにして事前に作成しておく．

まず，式 (5) において $\sigma = 1, \boldsymbol{q} = \mathbf{0}$ とした場合の確率分布 $p_{\mathrm{norm}}(\boldsymbol{x})$ を考える．

$$p_{\mathrm{norm}}(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}, \mathbf{I}) = \frac{1}{(2\pi)^{d/2}} \exp\left[-\frac{1}{2}\|\boldsymbol{x}\|^2\right] \qquad (6)$$

また，図 4 に示すように，原点から α の距離の点を中心点とする半径 δ の d 次元の超球 R が存在するとする．このとき，$p_{\mathrm{norm}}(\boldsymbol{x})$ を超球 R の領域で積分した値を，異なる α, δ の値に対して数値積分によって計算し，表 1 に示すような対応表を作成する．この表は，与えられた (α, δ) のペアに対して，対応する超球 R の領域での $p_{\mathrm{norm}}(\boldsymbol{x})$ の積分値を返す表である．

次に，事前に作成しておいたこの表をどのように使用して問合せ処理を行うかについて，2 次元の場合を例に説明する．

例 3　2 次元平面上に，式 (5) において $\sigma = 1, d = 2$ とした $p_q(\boldsymbol{x})|_{\sigma=1,d=2}$ によりその位置が表される問合せオブジェクト q と，問合せ対象オブジェクトが存在し，q が $PNNQ(q, \theta)$ を求める確率的最近傍問合せを発行したとする．このとき，ある問合せ対象オブジェクト o_i が問合せを満たすかどうかを効率的に評価するためには，すでに前節で述べたように，o_i のボロノイ領域の最小包囲円の領域で $p_q(\boldsymbol{x})|_{\sigma=1,d=2}$ を積分した値を求める必要がある．この積分値を素早く求めるために表を使用する．o_i のボロノイ領域の最小包囲円が，q の平均の位置 $\boldsymbol{q}$ からの距離が α_i であり，半径が δ_i の円であったとすると，2 次元の場合についてあらかじめ作成しておいた表 1 のような表から，(α_i, δ_i) のペアに一致するエントリを見つけ，対応する積分値を得る．この積分値が θ 未満である場合には，o_i は問合せを満たさないオブジェクトであるとして棄却することができる．一方で，表を引いて得た積分値が θ 以上であるからといって，必ず問合せを満たすとはいえないことに注意する．表を引いて得たのはあくまでも最小包囲円の領域での積分値であり，ボロノイ領域での積分値を大きめに近似した値として利用しているにすぎない．そのため，表を引いて得た積分値が θ 以上である場合には，ボロノイ領域での積分値を数値積分により計算することが必要になる．

与えられうる α, δ の値の組合せは膨大な数に上るため，すべての場合に対するエントリを表に登録しておくことは不可能である．そのため，例 3 の場合とは異なり，与えられた (α_i, δ_i) のペアにちょうど一致するようなエントリが表中に存在しないことがある．このような場合には，α の値が α_i 以下，かつ，δ の値が δ_i 以上であるような表中のエントリのうちで，できる限り積分値が小さいようなエントリを見つけて，対応する積分値を得る．つまりは，(α_i, δ_i) に対応する実際の積分値よりは大きいが，できる限りそれに近い値を返すようなエントリを見つける．実際の積分値より大きい値とするのは，実際の積分値より小さい値を積分値が θ 未満かどうかの判定に用いてしまうと，本来は問合せを満たすオブジェクトであるにも関わらず，積分値が θ 未満であるとして誤って棄却されてしまうということが起こりうるためである．実際の積分値より大きい値とすると，

逆に，表を引くだけで本来は棄却できるはずのオブジェクトが棄却されないということが起こりうるが，そのようなオブジェクトは，後にボロノイ領域での積分値を数値積分により求めた際にその値が θ 未満となるので，最終的な問合せ結果には含まれない．ただし，当然ながら，表中にちょうど一致するエントリが存在した場合よりも処理にコストがかかることになる．

作成した表は，さまざまな α, δ の組合せに対して，$\sigma = 1$ の場合の積分値を登録したものであるが，例 3 のような $\sigma = 1$ の場合だけではなく，異なる σ の場合についても，表を用いて積分値を求めることができる．一般の $p_q(\boldsymbol{x})$ に対する積分値を求めるためには，

$$\alpha' = \alpha/\sqrt{\sigma} \tag{7}$$

$$\delta' = \delta/\sqrt{\sigma} \tag{8}$$

として，(α', δ') により表を引けばよい．証明を以下に示す．

証明 1　簡単化のために $\boldsymbol{q} = \boldsymbol{0}$ とおく．このようにおいても以下の議論は一般性を失わない．$x_i/\sqrt{\sigma} = y_i$（$i = 1, 2, \ldots, d$）とおくと，以下のように変形できる．

$$\begin{aligned}\int_{\boldsymbol{x}\in R} p_q(\boldsymbol{x})d\boldsymbol{x} &= \int_{\boldsymbol{x}\in R} \frac{1}{(2\pi)^{d/2}\sigma^{d/2}} \exp\left[-\frac{1}{2\sigma}\|\boldsymbol{x}\|^2\right] d\boldsymbol{x} \\ &= \int_{\boldsymbol{y}\in R'} \frac{1}{(2\pi)^{d/2}\sigma^{d/2}} \exp\left[-\frac{1}{2}\|\boldsymbol{y}\|^2\right] \sigma^{d/2} d\boldsymbol{y} \\ &= \int_{\boldsymbol{y}\in R'} p_{\mathrm{norm}}(\boldsymbol{y})d\boldsymbol{y}\end{aligned}$$

ただし，R' は $\boldsymbol{x} \in R \Leftrightarrow \boldsymbol{y} \in R'$ を満たすような超球である．超球 R' の中心点が原点から $\alpha/\sqrt{\sigma}$ の距離にあり，半径が $\delta/\sqrt{\sigma}$ であることは容易に導ける．■

3.3　一般の場合

本節では，式 (3) の共分散行列が単位行列の定数倍とは限らない場合について考える．この場合の $p_q(\boldsymbol{x})$ の等確率面は楕円体の形状となり，前節で述べた，等確率面が超球の形状であるという単純な場合とは違って，$p_q(\boldsymbol{x})$ そのものを各ボロノイ領域の最小包囲円の領域で積分した値を表を引いて求めることは難しい．そこで，$p_q(\boldsymbol{x})$ の上限の関数である $p_q^\top(\boldsymbol{x})$ という関数を考えることにする．

共分散行列の逆行列 $\boldsymbol{\Sigma}^{-1}$ のスペクトル分解を

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{d} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^t \tag{9}$$

と定義する．$\lambda_i, \boldsymbol{v}_i$ はそれぞれ i 番目の固有値と固有ベクトルである（$i = 1, 2, \ldots, d$）．ただし，

$$\lambda^\top = \min\{\lambda_i\} \tag{10}$$

とする．共分散行列の固有値はすべて 0 より大であるので $\lambda^\top > 0$ が成り立つ．

ここで，行列 $\mathbf{M}^\top$ を

$$\mathbf{M}^\top = \lambda^\top \sum_{i=1}^{d} \boldsymbol{v}_i \boldsymbol{v}_i^t = \lambda^\top \mathbf{I} \tag{11}$$

と定義したとき，式 (3) の $\boldsymbol{\Sigma}^{-1}$ を $\mathbf{M}^\top$ で置き換えて得られる関数を $p_q^\top(\boldsymbol{x})$ と定義する．

$$p_q^\top(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{\lambda^\top}{2}\|\boldsymbol{x} - \boldsymbol{q}\|^2\right] \tag{12}$$

$p_q^\top(\boldsymbol{x})$ の等確率面は，d 次元の球の形状をしており，同じ確率について等確率面を描いたとき，式 (3) の確率密度関数の等確率面に外接する．なお，関数 $p_q^\top(\boldsymbol{x})$ は，積分しても 1 とはならないので，もはや確率密度関数ではない．$p_q^\top(\boldsymbol{x})$ には任意の $\boldsymbol{x}$ について以下が成り立つという性質がある．

$$p_q(\boldsymbol{x}) \leqq p_q^\top(\boldsymbol{x}) \tag{13}$$

また，$p_q^\top(\boldsymbol{x})$ は，このような性質を満たし，等確率面が球である関数のうちで，最良のものである．つまり，$p_q^\top(\boldsymbol{x})$ は $p_q(\boldsymbol{x})$ の上限を与えていることになる．$p_q^\top(\boldsymbol{x})$ のイメージを図 5 に示す．この図は各関数において同じ値（確率）が得られるような (x_i, x_j) 座標値の軌跡を示しており，ある特定の確率の値の軸の値についてスライスした状況にあたる．

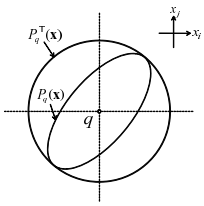

図 5　$p_q^\top(\boldsymbol{x})$ のイメージ

後述するが，この $p_q^\top(\boldsymbol{x})$ については，さまざまな半径および中心点を持つ円に対して，その円の領域での積分値を，前節で作成した表を引くことで簡単に求めることができる．そのため，問合せ処理において，最小包囲円の領域での $p_q(\boldsymbol{x})$ の積分値を求める代わりに，同じ領域での $p_q^\top(\boldsymbol{x})$ の積分値を求めることにする．$p_q^\top(\boldsymbol{x})$ には，任意の $\boldsymbol{x}$ について式 (13) が成り立つという性質があるため，同じ領域で積分した場合，その積分値が $p_q(\boldsymbol{x})$ の積分値を超えることはない．よって，最小包囲円の領域での積分値が θ 以上かどうかの判定の際に，$p_q(\boldsymbol{x})$ の積分値の代わりに $p_q^\top(\boldsymbol{x})$ の積分値を用いたとしても，問合せを満たすはずのオブジェクトが誤って棄却されてしまうことはない．

さまざまな半径および中心点を持つ d 次元の超球に対して，その超球の領域で $p_q^\top(\boldsymbol{x})$ を積分した値を，$p_{\mathrm{norm}}(\boldsymbol{x})$ の積分値を登録した表 1 のような表を引いて求めることが可能である．すでに前節で，$\boldsymbol{\Sigma} = \sigma\mathbf{I}$ の場合について，あらゆる σ に対して，表を引くことでさまざまな超球の領域での積分値を求められることを示した．本節で議論している一般の場合についても，

$$\sigma = \frac{1}{\lambda^\top} \tag{14}$$

として，前節同様，式 (7)(8) によって得られる (α', δ') により表を引けばよい．ただし，表を引いて得た値を $(\lambda^\top)^{d/2}|\boldsymbol{\Sigma}|^{1/2}$ で割る必要があることに注意する．証明を以下に示す．

証明 2　簡単化のために $\boldsymbol{q}=\mathbf{0}$ とおく．このようにおいても以下の議論は一般性を失わない．

$$\begin{aligned}&\int_{\boldsymbol{x}\in R} p_q^{\top}(\boldsymbol{x})d\boldsymbol{x}\\ &= \int_{\boldsymbol{x}\in R} \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{\lambda^{\top}}{2}\|\boldsymbol{x}\|^2\right] d\boldsymbol{x}\\ &= \frac{1}{(\lambda^{\top})^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \int_{\boldsymbol{x}\in R} \frac{(\lambda^{\top})^{d/2}}{(2\pi)^{d/2}} \exp\left[-\frac{\lambda^{\top}}{2}\|\boldsymbol{x}\|^2\right] d\boldsymbol{x}\end{aligned}$$

$\sigma = 1/\lambda^{\top}$ とすると，

$$= \frac{1}{(\lambda^{\top})^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \int_{\boldsymbol{x}\in R} \frac{1}{(2\pi)^{d/2}\sigma^{d/2}} \exp\left[-\frac{1}{2\sigma}\|\boldsymbol{x}\|^2\right] d\boldsymbol{x}$$

上式の積分部分が $\boldsymbol{\Sigma}=\sigma\mathbf{I}$ の場合の $\int_{\boldsymbol{x}\in R} p_q(\boldsymbol{x})d\boldsymbol{x}$ と一致している．■

以上の議論を踏まえ，問合せオブジェクトの位置の確率分布のパラメータ $(\boldsymbol{q},\boldsymbol{\Sigma})$ と確率の閾値 θ が与えられた際の確率的最近傍問合せの処理アルゴリズムを以下のアルゴリズム 1 に与える．ただし，以下の処理を事前に行っておくものとする．

1. 表 1 に示すような対応表を作成する．
2. ボロノイ図を作成し，各問合せ対象オブジェクトのボロノイ領域の最小包囲円をそれぞれ求める．
3. 各オブジェクトについて，自身の座標値，最小包囲円の情報（中心点，半径），対応するボロノイ領域に隣接するオブジェクトの座標値をファイルに記録する．

隣接するオブジェクトの座標値を記録しておくのは，数値積分のコストを削減するためである．各ボロノイ領域の形状が複雑なため，ボロノイ領域での積分値を計算する際に各ボロノイ領域を数式によって規定して数値積分を行うとコストが高くなる．そのため，サンプリング点と自身および近隣の各オブジェクトとの距離を計算することにより数値積分を行うことにする．

4. まとめと今後の課題

オブジェクトの位置が曖昧な位置として確率的に表現されている場合には，問合せ結果も確率的になるため，通常の最近傍問合せでは対応することができない．そこで，本研究では，通常の最近傍問合せに確率の閾値を設けた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した．本研究では，特に，その位置が正規分布の確率密度関数に従う問合せオブジェクトが，点オブジェクトを対象とした確率的最近傍問合せを発行するという状況を対象としている．

最も単純な問合せ処理方法は，各問合せ対象オブジェクトに対して，そのボロノイ領域で問合せオブジェクトの位置についての確率密度関数を積分し，積分値が閾値以上かどうかを調べるというものである．しかしながら，ボロノイ領域が複雑な形状をしていることに加えて，本研究では問合せオブジェクトの位置が正規分布に従うとしているため，そのような積分計算には高いコストがかかる．そこで，提案手法では，ボロノイ領域で積分を行う前に，ボロノイ領域の最小包囲円（球，超球）の

アルゴリズム 1 確率的最近傍問合せ

```
1: procedure PNNQ(q, Σ, θ)
2:     Σ の固有値分解により λ^⊤ を，また，|Σ| を求める
3:     各オブジェクトの情報が記録されたファイルをスキャンし，各オブジェクトについて，最小包囲円の中心点の q からの距離 α と最小包囲円の半径 δ から式 (14)(7)(8) によって得られる (α′, δ′) により表を引き，得られた値を (λ^⊤)^{d/2}|Σ|^{1/2} で割って最小包囲円の領域での積分値をそれぞれ求める
4:     積分値が θ 以上であるオブジェクトを，積分値の大きい順に候補オブジェクトの集合 C = {c_1, ..., c_m} に含める
5:         ▷ θ 未満のオブジェクトは明らかに問合せを満たさない
6:     sum ← 0
7:     for i = 1 to m do
8:         数値積分により，ボロノイ領域での積分値，すなわち，問合せを満たす確率 Pr(c_i) を得る
9:         sum ← sum + Pr(c_i)
10:        if Pr(c_i) ≧ θ then
11:            output c_i
12:        end if
13:        if sum > 1 − θ then
14:            return   ▷ 残りについては θ 以上になる可能性がない
15:        end if
16:    end for
17: end procedure
```

領域で積分を行うというアプローチをとる．円領域での積分値は事前に作成しておいた表を用いて簡単に求められるため，最小包囲円の領域での積分値はボロノイ領域での積分値よりも計算コストが低い．また，最小包囲円の領域での積分値はボロノイ領域での積分値よりも大きくなるという性質を持っている．よって，最小包囲円の領域での積分値が閾値未満のオブジェクトについては，明らかに問合せを満たさないオブジェクトであるとして，高いコストのかかるボロノイ領域での積分を行うことなく棄却することができる．なお，最小包囲円の領域での積分値を簡単に求めるための表は，標準正規分布の場合について一度作成しておくだけでよい．パラメータの異なるさまざまな正規分布の場合に対して，その表を用いて対応可能である．

今後は，問合せ処理の更なる効率化と詳細な実験に基づく評価に取り組みたい．

謝　辞

本研究の一部は，文部科学省科学研究費（19024037）の助成による．

文　献

[1] Jian Pei, Ming Hua, Yufei Tao, and Xuemin Lin. Query answering techniques on uncertain and probabilistic data: tutorial summary. In *Proc. ACM SIGMOD*, pp. 1357–1364, 2008.

[2] 石川佳治, 飯島裕一．曖昧な位置に基づく空間問合せ処理の効率化．電子情報通信学会第 19 回データ工学ワークショップ (DEWS2008), 2008.

[3] 浅野哲夫, 小保方幸次．LEDA で始める C/C++プログラミング：入門からコンピュータ・ジオメトリまで．サイエンス社, 2002.