曖昧な位置情報に基づく空間問合せの処理手法の効率化

飯島　裕一† 　石川　佳治††,†

† 名古屋大学大学院情報科学研究科　〒 464–8601 名古屋市千種区不老町
†† 名古屋大学情報基盤センター　〒 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　通常の空間問合せ処理手法の場合，問合せを行うオブジェクトの位置情報は正確であることが前提となるが，現実的にはセンサの誤差などにより曖昧な位置情報しか得られない場合が多い．そこで，空間問合せを行うオブジェクトの曖昧な位置が正規分布に従うという状況に焦点を合わせ，問合せ処理手法の効率化について検討する．
キーワード　空間データベース，空間問合せ，曖昧な位置，正規分布

Improving Efficiency of Processing Methods for Spatial Queries Based on Imprecise Location Information

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　Standard spatial query processing methods assume that location information of the query object is precise. However, in real life, systems usually cannot obtain precise location information due to sensor noise and estimation errors. Therefore we focus on the situation that the imprecise location of the query object is specified by an Gaussian distribution, and examine improving efficiency of query processing methods.
Key words　Spatial databases, spatial queries, imprecise locations, Gaussian distributions

1. はじめに

近年，位置情報を利用したアプリケーションにおいて，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きている．これは，現実世界のオブジェクトの位置は曖昧にしか得ることができない場合が多いことに起因している．例えば移動ロボットの分野では，センサ信号や移動履歴などを基にして統計的な手法を用いた自己位置推定が一般的に行われる [17] が，センサの測定誤差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．

本研究では，曖昧な位置を持つオブジェクトが，自らの位置から最も近く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する．具体的には，問合せを行うオブジェクト（以降，問合せオブジェクトと呼ぶ）の位置が**正規分布**で表現され，問合せの対象となるオブジェクト（以降，データオブジェクトと呼ぶ）が確定的な位置で表される**点データ**である状況を扱う．対象とする問合せとして，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**（probabilistic nearest neighbor query, PNNQ）を定義し，この問合せを効率的に処理するために 2 つの問合せ戦略を提案する．実験では，2 つの戦略にそれらのハイブリッド戦略を加えた 3 つの戦略について，様々なパラメータ設定の下で各戦略の比較を行う．

本稿ではまず，2 節で関連研究を紹介する．次に，3 節で確率的最近傍問合せを定義し，続く 4 節でその処理手法を提案する．5 節では評価実験について述べ，最後に 6 節でまとめを行う．

2. 関連研究

近年，位置の曖昧さを考慮した空間問合せ処理に関する多くの研究がなされてきたが，曖昧さを表現するためのモデルはさまざまである．例えば，[14] では曖昧なオブジェクトの位置が一様分布に従うことを前提としているが，[7,8,16] などでは任意の確率分布の使用を認めて一般性を高めている．一方で，本研究では位置の曖昧さが正規分布に従う状況を対象とする．移動オブジェクトデータベースの分野では，移動オブジェクトの位置の曖昧さを正規分布により表現するアイデアが [14] で提案されている．また，移動ロボットの分野では，センサや移動履歴を基にした自己位置推定にしばしば正規分布が使用される．

特に，正規分布に基づく確率過程を前提とした**カルマンフィルタ**は伝統的なアプローチである [17]．移動ロボットなどの応用を考えると，位置が正規分布に従うという想定には一般性があり，対象領域によっては十分な妥当性があるといえる．したがって，本研究では正規分布に特化した処理技術に焦点を合わせる．

曖昧さを考慮した空間問合せ処理に関する研究は各々の対象とする状況から以下の 3 種類に分類できる．

- データオブジェクトのみ曖昧 [9,10,14,16]
- 問合せオブジェクトのみ曖昧 [12]
- 両オブジェクトともに曖昧 [7,8,13]

本研究が対象とするのは問合せオブジェクトのみが曖昧であるという状況である．[12] は本研究グループによる研究であり，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の処理手法を提案している．本研究ではそのアイデアを一部導入しているが，対象とする問合せが最近傍問合せであるため，その特徴を考慮した改良や新しい技術が必要となる．最近傍問合せの処理にはデータオブジェクトに対する**ボロノイ図** [6] を用いるのが一般的であり，本研究でもボロノイ図を効果的に使用することで効率的な問合せ処理を実現する．データオブジェクトが曖昧である状況に対してボロノイ図を適用する場合には適切な拡張が必要となるが，本研究ではデータオブジェクトの位置が確定的である状況を扱うため通常のボロノイ図を利用できる．

3. 確率的最近傍問合せの定義

問合せを定義する前に，まず，問合せオブジェクトの位置を正規分布の確率密度関数によって確率的に定義する．

定義 3.1 (問合せオブジェクトの位置)
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 次元正規分布の確率密度関数により，

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

で表現されるとする．ただし，$\boldsymbol{\Sigma}$ は $d \times d$ の共分散行列，$\boldsymbol{q}$ は分布の平均，t はベクトルの転置を表す．

このように，問合せオブジェクトの位置が，曖昧なもの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決まることになる．そのような状況に対応するためには，各データオブジェクトに対する「q に最も近いオブジェクトとなる確率」を考慮して問合せ結果が定まるような最近傍問合せを定義する必要がある．ユークリッド距離に基づく通常の最近傍問合せを拡張する形で，以下にこれを定義する．

定義 3.2 (確率的最近傍問合せ)
問合せオブジェクト q 及び確率の閾値 θ $(0 < \theta < 1)$ が与えられたとき，q とのユークリッド距離がすべてのデータオブジェクトのうちで最小となる（q の最近傍オブジェクトとなる）確率が θ 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 $\mathcal{O}$ とするとき，$o \in \mathcal{O}$ が q の最近傍オブジェクトとなる確率 $\Pr_{NN}(q,o)$ は以下の式で表される．

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

これを用いて $PNNQ(q,\theta)$ は以下の式で表現できる．

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

問合せのパラメータとして与えられるのは，問合せオブジェクト q の情報と確率の閾値 θ である．q の情報とは，具体的には式 (1) に示した $p_q(\boldsymbol{x})$ の平均 $\boldsymbol{q}$ と共分散行列 $\boldsymbol{\Sigma}$ のことである．

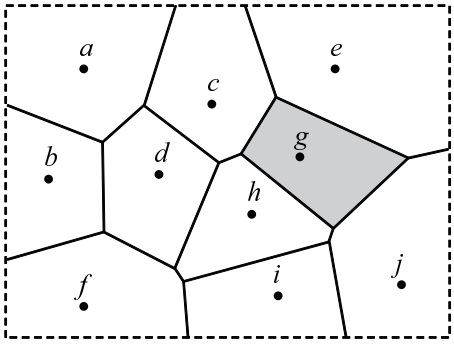

図 1　ボロノイ図

4. 提 案 手 法

4.1 基本的なアイデア

本手法ではボロノイ図と呼ばれる，空間中の複数の点に対して，どの点に一番近いかによって空間を分割した図を用いる．例として図 1 に点 a〜j に対するボロノイ図を示す．各点の勢力範囲は**ボロノイ領域**と呼ばれ，図 1 の陰影部分は g のボロノイ領域 V_g を示している．ボロノイ図の定義から，データオブジェクト o のボロノイ領域 V_o 内に q が位置するときに限って，o は q の最近傍オブジェクトとなる．したがって，o が q の最近傍オブジェクトとなる確率は q が V_o 内に位置する確率と言い換えられる．つまり，$\Pr_{NN}(q,o)$ は式 (1) に示した $p_q(\boldsymbol{x})$ を領域 V_o で積分することで計算できる．以上の事実を踏まえると，式 (2) に示した $\Pr_{NN}(q,o)$ の計算式は以下のとおりとなる．

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

本研究では $p_q(\boldsymbol{x})$ として正規分布の確率密度関数を用いているが，その積分は解析的には計算できないため，式 (4) の計算にはコストの高い数値積分が必要となる．その上，各ボロノイ領域の形状が複雑な多面体であることも計算コストを高める要因となる．したがって，すべてのデータオブジェクトに対して直接的に $\Pr_{NN}(\cdot)$ を求め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，候補オブジェクトと呼ぶ）に対してのみ $\Pr_{NN}(\cdot)$ を計算する．このアイデアに基づく問合せ戦略を 2 種類提案する．

4.2 矩形領域に基づく手法（RR 法）

本戦略では以下で定義する，**θ-領域** [12] という領域を用いてフィルタリングを行う．ただし，後述するとおり，実際には θ-領域を包含する矩形領域（rectilinear region）を用いることにしており，RR 法という名称はこれに由来している．

定義 4.1 (θ-領域)
$(\boldsymbol{x}-\boldsymbol{q})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{q}) \leqq r^2$ を満たす楕円体領域での $p_q(\boldsymbol{x})$ の積分を考える．与えられた θ $(0<\theta<1/2)$ に対し，積分値が $1-2\theta$ になるような r の値を r_θ とする．式で表現すると以下のとおりである．

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

r_θ により以下の式で定まる楕円体領域を θ-領域と呼ぶ．

$$(\boldsymbol{x}-\boldsymbol{q})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{q}) \leqq r_\theta^2 \tag{6}$$

θ-領域は問合せ時のパラメータに依存するため，その導出は問合せ時に動的に行う必要がある．単純な方法として，問合せ時に様々な r の値に対して対応する楕円体領域での $p_q(\boldsymbol{x})$ の積分値を数値積分によって計算し，その値が $1-2\theta$ となるような $r=r_\theta$ を見つけるという方法が考えられるが，計算コストの面で現実的ではない．そこで，楕円体領域での積分を球領域での積分に変換する．まず，式 (1) において $\boldsymbol{q}=\boldsymbol{0}, \boldsymbol{\Sigma}=\mathbf{I}$ とした，**標準正規分布**の確率密度関数

$$p_{\mathrm{norm}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}}\exp\left[-\frac{1}{2}\|\boldsymbol{x}\|^2\right] \tag{7}$$

を考える．これを用いて以下の性質を導出することができる．証明は [12] を参照されたい．

性質 4.1
原点を中心とした半径 r の球領域 $\|\boldsymbol{x}\|^2 \leqq r^2$ での $p_{\mathrm{norm}}(\boldsymbol{x})$ の積分を考える．与えられた θ に対し，積分値が $1-2\theta$ になるような半径を $\tilde{r}_\theta$ と定義する．式で表現すると以下のとおりである．

$$\int_{\|\boldsymbol{x}\|^2\leqq\tilde{r}_\theta^2} p_{\mathrm{norm}}(\boldsymbol{x})d\boldsymbol{x} = 1-2\theta \tag{8}$$

このとき，与えられた θ に対して以下の式が成り立つ．

$$r_\theta = \tilde{r}_\theta \tag{9}$$

この性質は，与えられた θ に対して，式 (8) に基づいて $\tilde{r}_\theta$ を計算すれば，その値がそのまま θ-領域を定める r_θ になっているということを示している．しかしながら，$p_{\mathrm{norm}}(\boldsymbol{x})$ の積分値は解析的に求めることができないため，θ から直接 r_θ を計算することはできない．そこで逆に，適当な半径の値を選んでその半径を持つ球領域での $p_{\mathrm{norm}}(\boldsymbol{x})$ の積分値を数値積分によって計算するということを，様々な半径の値に対して行うことで，積分値から得られる θ とそのときの半径 r_θ の対応表を事前に作成しておくことにする．この表を引くことで与えられた θ に対応する r_θ を素早く得ることが可能となる．式 (8) の計算は，θ 以外の問合せのパラメータである $\boldsymbol{q}$ 及び $\boldsymbol{\Sigma}$ に依存しないため，このような方法をとることができる．このようなアイデアは [16] でも導入されており，表は *U-catalog* と呼ばれている．

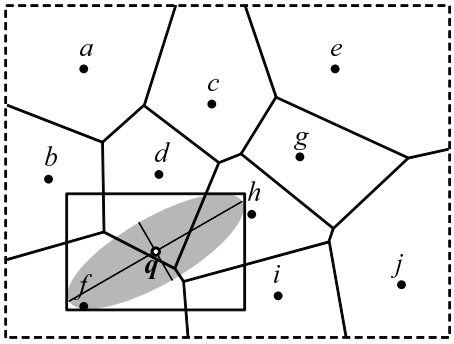

図 2　矩形領域とボロノイ領域

アルゴリズム 1 RR 法に基づく確率的最近傍問合せ

```
1: procedure PNNQ-RR(q, Σ, θ)
2:     C ← ∅, sum ← 0
3:     r_θ ← lookup(θ)                    ▷ U-catalog から r_θ を得る
4:     θ-領域を包含する矩形領域を導出
5:     ボロノイ領域が矩形領域と重なりを持つオブジェクトを検索して C に挿入
6:     foreach o ∈ C do
7:         Pr_NN(q, o) ← ∫_{x∈V_o} p_q(x)dx    ▷ 数値積分による
8:         sum ← sum + Pr_NN(q, o)
9:         if Pr_NN(q, o) ≧ θ then
10:            output o
11:        end if
12:        if sum > 1 − θ then
13:            return
14:        end if
15:    end for
16: end procedure
```

図 2 を用いて本戦略のアイデアを説明する．図の陰影部分で示された楕円体領域が θ-領域であるが，これを直接フィルタリングに利用することは難しいため，これを包含する，座標軸に平行な矩形領域を考えることにする(注1)．ここで，ボロノイ領域が矩形領域と重なりを持たないオブジェクト，すなわち a, c, e, g, j を解の候補から除外することができる．理由は以下のとおりである．まず，θ-領域の定義から，矩形領域の外側の領域全体での $p_q(\boldsymbol{x})$ の積分値は $1-(1-2\theta)=2\theta$ 未満である．また，$p_q(\boldsymbol{x})$ は分布の平均 $\boldsymbol{q}$ について点対称な分布であるため，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，矩形領域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．すなわち，ボロノイ領域が矩形領域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として除去できる．

本戦略のアルゴリズムをアルゴリズム 1 に示す．ボロノイ領域が矩形領域と重なりを持つオブジェクトを検索して候補オブ

(注1)：矩形領域は r_θ と $\boldsymbol{\Sigma}$ から導出できる．詳細は [12] を参照されたい．

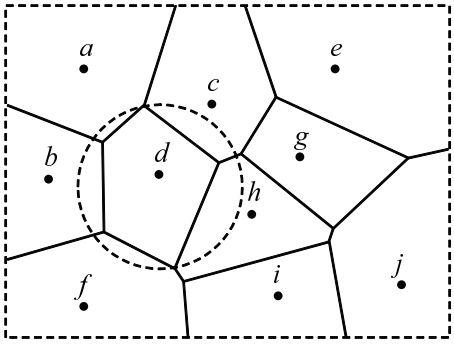

図 3　最小包含球

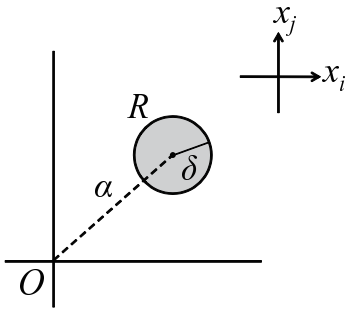

図 4　球　　R

α	δ	$\pi(\alpha,\delta)$
0.0	0.1	$\cdots$
0.0	0.2	$\cdots$
$\vdots$	$\vdots$	$\vdots$
1.0	0.1	$\cdots$
1.0	0.2	$\cdots$
$\vdots$	$\vdots$	$\vdots$

図 5　BF 法の U-catalog

ジェクトとした後，すべての候補オブジェクトに対して，数値積分により $\Pr_{NN}(\cdot)$ を求め，θ 以上であれば出力するという流れで処理を行う．ただし，U-catalog の作成と各ボロノイ領域の頂点の座標のファイルへの記録を事前に行っておくものとする．5 行目でボロノイ領域が矩形領域と重なりを持つオブジェクトを検索する必要があるが，各ボロノイ領域は複雑な形状をとるため，そのようなオブジェクトだけを正確に検索する処理はコストが高い．解決策としては，各ボロノイ領域に対しても，それらを包含する，座標軸に平行な矩形領域を求め，この矩形領域が θ-領域を包含する矩形領域と重なりを持つオブジェクトを候補オブジェクトとすることが考えられる．これにより，若干候補オブジェクトの個数が増加することになるが，フィルタリングに要する時間を減らすことができる．予備実験の結果，候補オブジェクトの増加による確率計算の処理時間の増加分をフィルタリングの処理時間の減少分が上回る見込みが濃厚であったため，5 節の実験ではこの方法に基づいて実装を行った．12 行目の条件が満たされると，その時点で残っている候補については $\Pr_{NN}(\cdot)$ が θ 以上である可能性がなくなるため，処理を終了できる．

4.3　上限関数に基づく手法（BF 法）

本戦略では，各データオブジェクトに対して $\Pr_{NN}(\cdot)$ の上限値を求めることによりフィルタリングを行う．上限値の計算はボロノイ領域の**最小包含球**（smallest enclosing sphere, SES）を利用して行う．例としてボロノイ領域 V_d の最小包含球を図 3 に示す．最小包含球の領域で $p_q(\boldsymbol{x})$ を積分すると，その値は $\Pr_{NN}(\cdot)$ の上限値とみなすことができる．球領域での積分値は事前に表を作成しておくことで簡単に求められるため，最小包含球による上限値の計算は高速なフィルタリング処理の実現に有効である．このことについて，以下で説明を行う．始めに式 (1) の共分散行列 $\boldsymbol{\Sigma}$ が単位行列であるという単純な場合を考え，次にアイデアを一般の場合に拡張する．詳細は後述するが，一般の場合には $p_q(\boldsymbol{x})$ の代わりに，その**上限関数**（upper-bounding function）を利用する．BF 法という名称はこれに由来している．

4.3.1　$\boldsymbol{\Sigma}=\mathbf{I}$ の場合

本節では $\boldsymbol{\Sigma}$ が単位行列である場合について考える．この場合の $p_q(\boldsymbol{x})$ は $p_{\mathrm{norm}}(\boldsymbol{x})$ を $\boldsymbol{q}$ が中心となるように平行移動したものに等しい．

最小包含球の半径や中心の座標はオブジェクトごとに様々であるため，異なる最小包含球に対して，その領域での $p_q(\boldsymbol{x})$ の積分値を素早く導出できるように対応表を事前に作成しておく．表の作成にあたり，図 4 に示すような，原点から距離 α の点を中心とする半径 δ の d 次元の球 R を考える．このとき，$p_{\mathrm{norm}}(\boldsymbol{x})$ を R の領域で積分した値を以下の式で表す．

$$\pi(\alpha,\delta)=\int_{\boldsymbol{x}\in R} p_{\mathrm{norm}}(\boldsymbol{x})d\boldsymbol{x} \tag{10}$$

$p_{\mathrm{norm}}(\boldsymbol{x})$ の等確率面は球形であり，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ため，このような表記を用いることができる．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 5 に示すような表を作成する．この表も RR 法で用いた表と同様に U-catalog と呼ぶ．U-catalog は (α,δ) のペアを与えると対応する積分値を返す．

次に，U-catalog の使用方法を説明する．本戦略では，ボロノイ領域 V_o の最小包含球を SES_o としたときの $\int_{\boldsymbol{x}\in SES_o} p_q(\boldsymbol{x})d\boldsymbol{x}$ の値を，$\Pr_{NN}(q,o)$ の上限値として用いる．この値は，$\boldsymbol{q}$ 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) に一致するエントリを U-catalog から検索すれば簡単に得られる．得られた値が θ 以下である場合には o を棄却できる．一方，この値はあくまでも $\Pr_{NN}(q,o)$ の上限値であるため，値が θ より大きいからといって o が解になるとは限らない．そのため，U-catalog を引いて得られた値が θ より大きいオブジェクトは候補オブジェクトとして残す．

4.3.2　一般の場合

本節では $\boldsymbol{\Sigma}$ が任意である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をとるため，単純に (α_o,δ_o) のペアによって任意の球領域での積分値を表すことは不可能であり，表を用いて積分値を求めるわけにはいかない．そこで，以下に式を示す，$p_q(\boldsymbol{x})$ の上限関数 $p_q^{\top}(\boldsymbol{x})$ を導入する．

$$p_q^{\top}(\boldsymbol{x})=\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{\lambda^{\top}}{2}\|\boldsymbol{x}-\boldsymbol{q}\|^2\right] \tag{11}$$

ただし，$\lambda^{\top}$ は $\boldsymbol{\Sigma}^{-1}$ の固有値のうちで最小のものである．

$p_q^{\top}(\boldsymbol{x})$ の等確率面は球形である．ただし，空間全体での積分値が 1 とはならないため，この関数は厳密には確率密度関数ではない．$p_q^{\top}(\boldsymbol{x})$ は以下の性質を持つ．

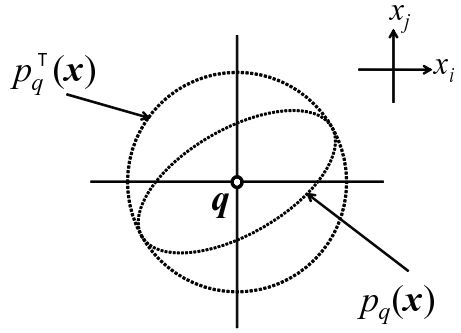

図 6　同じ確率に対する $p_q(\boldsymbol{x})$ と $p_q^\top(\boldsymbol{x})$ の等確率面

性質 4.2

任意の $\boldsymbol{x}$ に対して，以下の式が成り立つ．

$$p_q(\boldsymbol{x}) \leqq p_q^\top(\boldsymbol{x}) \tag{12}$$

同じ確率に対する $p_q(\boldsymbol{x})$ と $p_q^\top(\boldsymbol{x})$ の等確率面を図 6 に示す．$p_q^\top(\boldsymbol{x})$ の等確率面は $p_q(\boldsymbol{x})$ の等確率面に外接する．つまり $p_q^\top(\boldsymbol{x})$ は，性質 4.2 を満たし，等確率面が球形の関数のうちで最良のものであり，$p_q(\boldsymbol{x})$ の上限を与えているといえる．

$p_q^\top(\boldsymbol{x})$ については等確率面が球形であるため，任意の球領域での積分値を表により求めることが可能である．その上，表は 4.3.1 節で説明した $\boldsymbol{\Sigma} = \mathbf{I}$ の場合についての U-catalog をそのまま使用することができる．具体的には，$(\alpha_o\sqrt{\lambda^\top}, \delta_o\sqrt{\lambda^\top})$ に一致するエントリを U-catalog から検索し，得られた $\pi(\alpha_o\sqrt{\lambda^\top}, \delta_o\sqrt{\lambda^\top})$ を $(\lambda^\top)^{d/2}|\boldsymbol{\Sigma}|^{1/2}$ で割ることで求められる．証明は [12] を参照されたい．性質 4.2 より，同じ領域で積分した場合に，$p_q^\top(\boldsymbol{x})$ の積分値が $p_q(\boldsymbol{x})$ のそれを下回ることはないため，最小包含球領域での $p_q^\top(\boldsymbol{x})$ の積分値が θ 以下であるオブジェクトは $\mathrm{Pr}_{NN}(\cdot)$ が θ 以上になることはないとして棄却できる．

本戦略のアルゴリズムをアルゴリズム 2 に示す．ただし，U-catalog の作成と各ボロノイ領域の頂点の座標，最小包含球の中心点及び半径のファイルへの記録を事前に行っておくものとする．12 行目のソートにより，最小包含球領域での $p_q^\top(\boldsymbol{x})$ の積分値が大きいオブジェクトから順に $\mathrm{Pr}_{NN}(\cdot)$ を計算できる．この値が大きいオブジェクトはボロノイ領域での $p_q(\boldsymbol{x})$ の積分値，すなわち $\mathrm{Pr}_{NN}(\cdot)$ も大きいと考えられるため，順序を考慮しない場合よりも早く処理を終了できる可能性が高い．

アルゴリズム 2 BF 法に基づく確率的最近傍問合せ

```
 1: procedure PNNQ-BF(q, Σ, θ)
 2:     C ← ∅, sum ← 0
 3:     λ^⊤ 及び |Σ| を Σ から計算
 4:     foreach o ∈ O do
 5:         q から SES_o の中心までの距離 α_o を計算
 6:         π(α_o√λ^⊤, δ_o√λ^⊤) ← lookup(α_o√λ^⊤, δ_o√λ^⊤)
                    ▷ U-catalog から π(α_o√λ^⊤, δ_o√λ^⊤) を得る
 7:         IV_SES_o ← π(α_o√λ^⊤, δ_o√λ^⊤)/(λ^⊤)^{d/2}|Σ|^{1/2}
                                  ▷ SES_o での p_q^⊤(x) の積分値
 8:         if IV_SES_o > θ then
 9:             C ← C ∪ {o}
10:         end if
11:     end for
12:     C 中のオブジェクトを IV_SES_o の降順でソート
13:     foreach o ∈ C do                          ▷ 先頭から順に
14:         Pr_NN(q, o) ← ∫_{x∈V_o} p_q(x)dx      ▷ 数値積分による
15:         sum ← sum + Pr_NN(q, o)
16:         if Pr_NN(q, o) ≧ θ then
17:             output o
18:         end if
19:         if sum > 1 − θ then
20:             return
21:         end if
22:     end for
23: end procedure
```

5. 実験に基づく提案手法の評価

5.1 実験方法

使用したデータは米国カリフォルニア州ロングビーチにおける道路の線分データ [5] から各線分の中点を抽出して作成されたデータである．データ数は 50,501 で，各点は $[0, 1000]^2$ の 2 次元空間上に位置するように正規化されている．各点をデータオブジェクトとして，2 つの問合せ戦略にそれらのハイブリッド戦略を加えた 3 つの戦略を対象に，$PNNQ(q, \theta)$ に対する性能を評価した．

ハイブリッド戦略は，RR 法のフィルタリングと BF 法のフィルタリングを組み合わせたフィルタリングを行う戦略である．具体的には，始めに RR 法のフィルタリングを行った後，残ったオブジェクトに対して BF 法のフィルタリングを行う．そのため，得られる候補オブジェクトの集合は，RR 法のフィルタリングによって得られる候補オブジェクトの集合と BF 法のフィルタリングによって得られる候補オブジェクトの集合の積集合になる．RR 法のフィルタリングを行ってから BF 法のフィルタリングを行うという順序には理由があるが，これについては実験の結果を踏まえながら 5.2.1 節で述べる．

式 (1) の共分散行列 $\boldsymbol{\Sigma}$ の設定は以下を標準とした．

$$\boldsymbol{\Sigma} = \gamma \begin{bmatrix} 7 & 2\sqrt{3} \\ 2\sqrt{3} & 3 \end{bmatrix}$$

これにより，$p_q(\boldsymbol{x})$ の等確率線の形状は長軸と短軸の比が 3 : 1 で傾き 30° の楕円となる．係数 γ は分布の曖昧さの程度に対応する．この実験では $\gamma = 10, \theta = 0.03$ を標準の設定とし，そこから値を変動させることで γ 及び θ が各戦略の性能に与える影響を調べた．また，$\boldsymbol{\Sigma}$ を変えることで $p_q(\boldsymbol{x})$ の等確率線の形状が異なる場合についても評価を行った．

各戦略ごとに，$\boldsymbol{q}$ の異なる 100 回の問合せ処理を行い，その平均応答時間を性能の評価基準に用いた．ただし，応答時間は $\boldsymbol{q}$ によって大きく影響を受けるため，100 回分の問合せの $\boldsymbol{q}$ は各戦略ごとにすべて同じものを使用した．事前に作成した U-catalog のエントリ数は，RR 法の U-catalog が 607，BF 法の U-catalog が 31,008 であった．

今回の実験に用いた問合せ処理プログラムでは，ボロノイ領域の計算に Qhull [3] を，最小包含球（今回の実験は 2 次元の

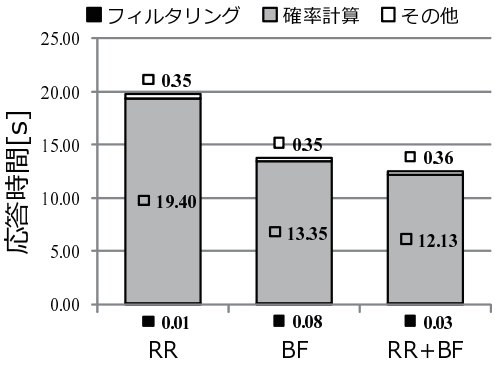

図 7　応答時間（$\gamma = 10, \theta = 0.03$）

表 1　候補オブジェクト数（$\gamma = 10, \theta = 0.03$）

RR	BF	RR+BF	解
101.5	62.3	55.0	5.4

場合を対象としているので，厳密には最小包含円）の計算に Miniball [11] をそれぞれ使用した．また，数値積分処理には RANDLIB [4] という C 言語の乱数生成ライブラリを用いた．具体的には，RANDLIB により正規分布の確率密度関数に従って大量の乱数を生成し，各乱数がボロノイ領域内に位置しているかどうかを LEDA [2] で提供されている関数を利用して調べた．LEDA はグラフ理論や幾何学計算などの分野における効率的なデータ構造とアルゴリズムを提供する C++のクラスライブラリである．ボロノイ領域内に位置していた乱数の個数の比率が求める確率の推定値に相当している．この手法は**重点サンプリング法** [15] と呼ばれ，モンテカルロ法の一種であるが，通常のモンテカルロ法による計算より高速である．今回の実験では，標準の設定として，1 回の積分計算に対して 1,000,000 個の乱数を発生させて積分値を求めるように設定した．BF 法で必要となる固有値や行列式の計算には，科学技術計算用の C 言語のライブラリである GNU Scientific Library [1] を用いた．

実験用プログラムの開発には C++を用いた．実験に使用したマシンの CPU は Intel Core 2 Duo E8500 (3.16GHz)，メモリは 4GB，OS は Fedora 12 である．

5.2 実験結果

5.2.1 標準の設定の場合

標準の設定（$\gamma = 10, \theta = 0.03$）における各戦略の応答時間を図 7 に，候補オブジェクト及び解オブジェクトの個数を表 1 に示す．また，ある問合せにおける各戦略の候補オブジェクトを図 8，9，10 に示す．分布の平均 $\boldsymbol{q}$ は図の中心に位置している．太い線でボロノイ領域が縁取られているオブジェクトが候補オブジェクトであり，ボロノイ領域が黒く塗りつぶされているオブジェクトが解オブジェクトである．図 8，10 における矩形は，RR 法のフィルタリングに用いる矩形領域を示している．図 7 に示されるように，各戦略とも処理時間のほとんどは $\Pr_{NN}(\cdot)$ の計算に費やされていた．これは予想通りの結果であり，フィルタリングによって $\Pr_{NN}(\cdot)$ の計算を必要とするオブジェクトの削減を図るという本手法のアプローチの正しさが確認された．

表 1 より，RR 法及び BF 法をそれぞれ単独で用いた場合の候補オブジェクト数は，RR 法が 101.5 個，BF 法が 62.3 個であるが，ハイブリッド戦略では 55.0 個に減っていることがわかる．これは，ハイブリッド戦略では RR 法と BF 法に共通の候補オブジェクトのみを候補とするため，一方の戦略では候補になるが他方の戦略では棄却されるようなオブジェクトがすべて棄却されるからである．

ハイブリッド戦略では RR 法のフィルタリングを行った後に BF 法のフィルタリングを行うが，図 7 を見ると，フィルタリングに要した時間は 0.03 秒であり，単純に，RR 法でフィルタリングに要した時間 0.01 秒と BF 法でフィルタリングに要した時間 0.08 秒の和にはなっていないことがわかる．この理由は，ハイブリッド戦略では RR 法のフィルタリングによって残ったオブジェクトに対してのみ BF 法のフィルタリングを適用することになるため，BF 法を単独で用いる場合よりも BF 法のフィルタリングに必要な時間を減らすことができるからである．フィルタリングの適用順を入れ替えた場合，得られる候補オブジェクトの集合は変わらないが，少なくとも BF 法を単独で用いる場合の 0.08 秒はフィルタリングに必要となってしまう．そのため，より短時間でフィルタリング処理が可能な「RR 法から BF 法」という順序を採用している．

この実験では 5.4 個という少数の解オブジェクトを返すのにハイブリッド戦略の場合でも約 13 秒を要しており，若干処理時間が長いように思われる．処理時間の短縮に最も効果的な方法は，その大部分を占めている確率計算に要する時間の短縮であり，数値積分に使用する乱数の個数（以下，サンプル数）を減らせばこれを達成できることは明らかである．つまり，今回の実験ではサンプル数を 1,000,000 個に設定しているが，例えばこれを 100,000 個に減らすことで，計算の精度は悪化するものの確率計算に要する時間をおよそ 1/10 にまで短縮できる．詳細は省略するが，実際にサンプル数を減らした場合について実験を行った結果から，多くの現実的な状況ではサンプル数を減らすことができる可能性が高いと考えている．ただし，サンプル数の設定は，計算時間と計算精度のトレードオフを考慮しながら，適用するアプリケーションの要件やユーザの設定に応じて適切に決定する必要があり，一概には論じられない．

5.2.2 γ を変動させた場合

γ を変動させた場合の各戦略の応答時間を図 11 に，候補オブジェクト及び解オブジェクトの個数を表 2 に示す．γ の大小は曖昧さの程度，具体的には $p_q(\boldsymbol{x})$ の等確率線の大小に相当している．そのため，γ の増大にしたがって θ-領域も広がることになり，図 11 に示したとおり，RR 法の性能は悪化する．一方，BF 法では γ の増大の影響を受けていない．これは，図 8，9 からわかるように，RR 法では $\boldsymbol{q}$ に近いオブジェクトはすべて候補となるが，BF 法では各オブジェクトに対して積分値の上限値を見積もることで候補かどうかを判定するため，そのようなオブジェクトでも棄却できる可能性があるということに関係がある．$p_q(\boldsymbol{x})$ の等確率線が大きくなるということは，それだけなだらかに広がった分布になるということであり，分布の平均 $\boldsymbol{q}$ から比較的近くに位置するオブジェクトについては，ボロノ

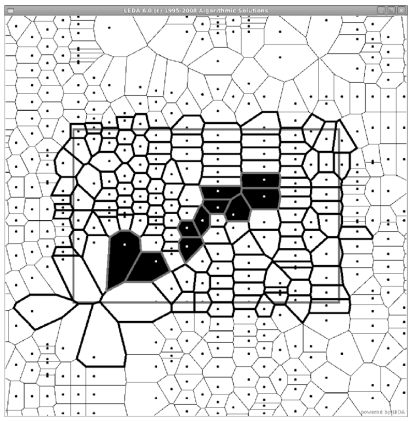

図 8　RR 法における候補オブジェクト

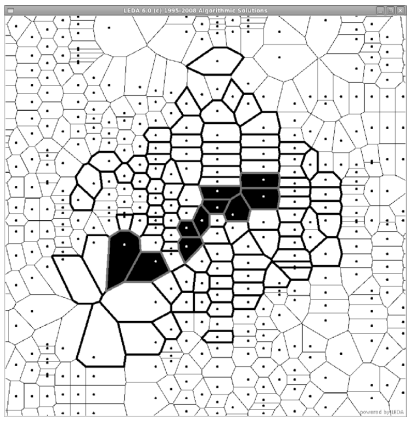

図 9　BF 法における候補オブジェクト

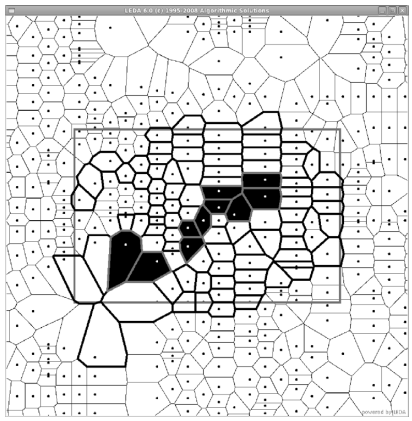

図 10　RR+BF 法における候補オブジェクト

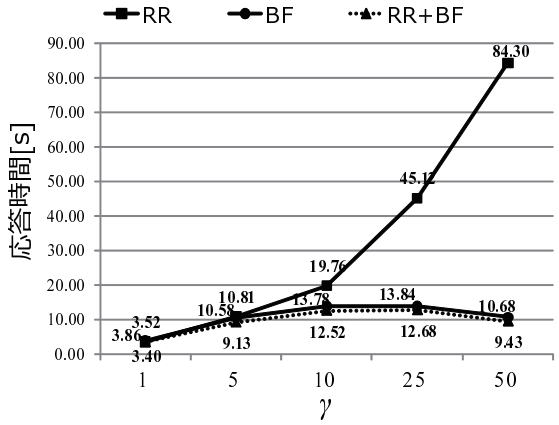

図 11　応答時間（γ を変動）

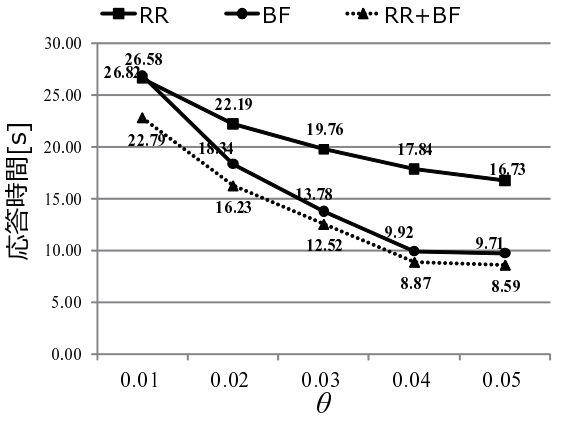

図 12　応答時間（θ を変動）

表 2　候補オブジェクト数及び解オブジェクト数（γ を変動）

	RR	BF	RR+BF	解
$\gamma = 1$	15.3	20.9	14.4	4.9
$\gamma = 5$	53.8	48.8	40.2	7.1
$\gamma = 10$	101.5	62.3	55.0	5.4
$\gamma = 25$	232.0	62.5	57.0	3.3
$\gamma = 50$	431.4	46.9	41.2	2.6

表 3　候補オブジェクト数及び解オブジェクト数（θ を変動）

	RR	BF	RR+BF	解
$\theta = 0.01$	137.8	125.7	105.5	18.9
$\theta = 0.02$	115.7	84.0	73.0	9.7
$\theta = 0.03$	101.5	62.3	55.0	5.4
$\theta = 0.04$	93.1	44.1	38.3	3.6
$\theta = 0.05$	85.8	42.5	36.9	2.5

イ領域での積分値が減少し，解でなくなる可能性が高まる．つまり，BF 法では γ の増大によって候補となるオブジェクトも棄却できるオブジェクトも両方存在するが，RR 法では候補となるオブジェクトしか存在しないのである．

5.2.3　θ を変動させた場合

θ を変動させた場合の各戦略の応答時間を図 12 に，候補オブジェクト及び解オブジェクトの個数を表 3 に示す．θ の増大に伴って BF 法の RR 法に対する優位性が高くなっていることがわかる．

5.2.4　$p_q(\boldsymbol{x})$ の等確率線の形状を変動させた場合

$p_q(\boldsymbol{x})$ の等確率線の形状が円の場合及び細い楕円（長軸と短軸の比が 9 : 1 で傾き 30° の楕円）の場合の応答時間をそれぞれ図 13，14 に，候補オブジェクト及び解オブジェクトの個数をそれぞれ表 4，5 に示す．また，ある問合せにおけるハイブリッド戦略の候補オブジェクトをそれぞれ図 15，16 に示す．円の場合には BF 法の方が性能が良いが，細い楕円の場合には優劣関係が逆転していることがわかる．この理由は以下のとおりである．BF 法では最小包含球の領域での積分値を求めるにあたって，$p_q(\boldsymbol{x})$ の代わりに $p_q^{\top}(\boldsymbol{x})$ の積分値を求めるが，図 6 に示したとおり，$p_q^{\top}(\boldsymbol{x})$ の等確率面は $p_q(\boldsymbol{x})$ の等確率面の外接球となる．したがって，$p_q(\boldsymbol{x})$ の等確率面の楕円体の形状が球に近い場合には，$p_q^{\top}(\boldsymbol{x})$ の積分値が $p_q(\boldsymbol{x})$ の積分値に近づくため，BF 法のフィルタリングの効率が良くなり，逆に楕円体の形状が細い場合には効率が悪くなるのである．

6. ま と め

本研究では，位置が正規分布によって曖昧な位置情報で表現されている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確率計算の対象から除外するとい

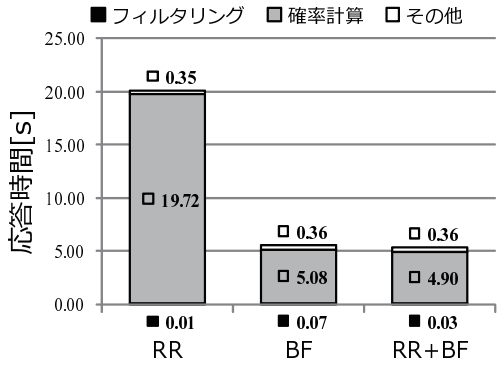

図 13　応答時間（円）

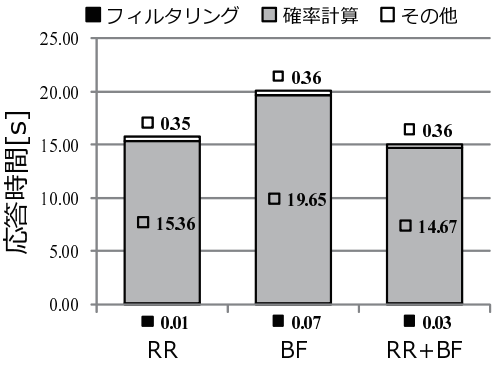

図 14　応答時間（細い楕円）

表 4　候補オブジェクト数及び解オブジェクト数（円）

RR	BF	RR+BF	解
100.6	21.8	20.8	4.1

表 5　候補オブジェクト数及び解オブジェクト数（細い楕円）

RR	BF	RR+BF	解
82.0	116.6	75.5	7.4

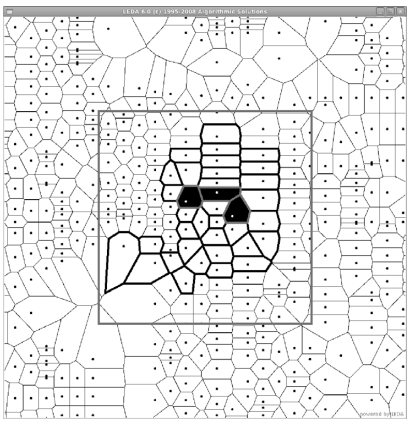

図 15　RR+BF 法における候補オブジェクト（円）

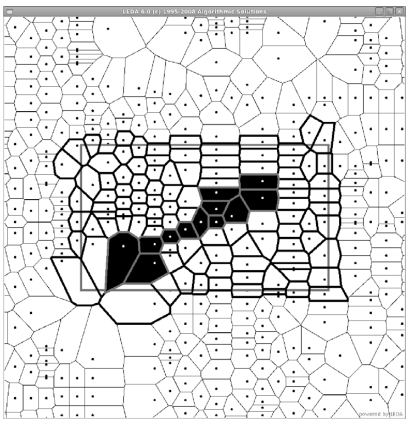

図 16　RR+BF 法における候補オブジェクト（細い楕円）

うアプローチに基づいて，θ-領域を包含する矩形領域に基づく戦略と，最小包含球と上限関数により確率の上限値を求める戦略を提案した．実験の結果，基本的には後者の方が性能が良いことがわかり，特に，位置の曖昧さが大きい，閾値が高い，正規分布の等確率面の楕円体の形状が球形に近い，という状況ではその傾向が顕著であった．ただし，実用性の観点からすると，2 つの戦略の組合せによって問合せ処理を行うのが良い．

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．

文　　献

[1] GNU Scientific Library. http://www.gnu.org/software/gsl/.
[2] LEDA. http://www.algorithmic-solutions.com/leda/.
[3] Qhull. http://www.qhull.org/.
[4] RANDLIB. http://biostatistics.mdanderson.org/SoftwareDownload/.
[5] TIGER. http://tiger.census.gov/.
[6] F. Aurenhammer. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3):345–405, 1991.
[7] G. Beskales, M. A. Soliman, and I. F. Ilyas. Efficient search for the top-k probable nearest neighbors in uncertain databases. In *Proc. VLDB*, pp. 326–339, 2008.
[8] J. Chen and R. Cheng. Efficient evaluation of imprecise location-dependent queries. In *Proc. ICDE*, pp. 586–595, 2007.
[9] R. Cheng, J. Chen, M. Mokbel, and C.-Y. Chow. Probabilistic verifiers: evaluating constrained nearest-neighbor queries over uncertain data. In *Proc. ICDE*, pp. 973–982, 2008.
[10] R. Cheng, D. V. Kalashnikov, and S. Prabhakar. Querying imprecise data in moving object environments. *IEEE TKDE*, 16(9):1112–1127, 2004.
[11] B. Gärtner. Miniball. http://www.inf.ethz.ch/personal/gaertner/miniball.html.
[12] Y. Ishikawa, Y. Iijima, and J. X. Yu. Spatial range querying for Gaussian-based imprecise query objects. In *Proc. ICDE*, pp. 676–687, 2009.
[13] H.-P. Kriegel, P. Kunath, and M. Renz. Probabilistic nearest-neighbor query on uncertain objects. In *Proc. DASFAA*, pp. 337–348, 2007.
[14] D. Pfoser and C. S. Jensen. Capturing the uncertainty of moving-object representations. In *Proc. 6th Intl. Symp. on Advances in Spatial Databases (SSD'99)*, pp. 111–131, 1999.
[15] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes: the art of scientific computing*. Cambridge University Press, 3rd edition, 2007.
[16] Y. Tao, X. Xiao, and R. Cheng. Range search on multidimensional uncertain data. *ACM TODS*, 32(3):15, 2007.
[17] S. Thrun, W. Burgard, and D. Fox. *Probabilistic robotics*. The MIT Press, 2005.