学位報告 4

報告番号	※甲　　第　　号

主　論　文　の　要　旨

論文題目　Query Processing over Probabilistic Data with Gaussian Distributions
（ガウス分布に基づく確率的データに対する問合せ処理）

氏　　名　　董　婷　婷

論 文 内 容 の 要 旨

センサネットワークや位置に基づくサービス，モニタリングと監視などの様々な実世界アプリケーションにおける不確実なデータの管理要求の増加により，不確実なデータの管理問題は研究者の大きな注目を集めている．不確実性が生じる原因は様々である．センサの測定誤差とノイズ，プライバシー保護のための機密を含むレコードの変換，予測モデルの出力結果における確信度の情報などがあげられる．

不確実データの管理は，不確実なデータのモデリングと表現，問合せ，索引付けに関わる．不確実データの管理にはコストのかかる確率計算が必要であるため，不確実なデータベースに対する問合せ処理には，伝統的なデータベースに対する問合せ処理と比較して，より多くの課題が存在する．そのため，不確実なデータの管理のための効率的な解決策が必要となる．本論文では，確率的モデルを用いて，データベースに蓄積された不確実なオブジェクトをガウス分布（Gaussian distribution）で表現する．ガウス分布は統計学やパターン認識，機械学習などの分野でよく用いられる代表的な確率分布である．ガウス分布に基づく確率的データに対して，以下のような三種類の問合せを考える．

第一に，不確実なデータの管理において重要な問合せである確率的範囲問合せ（probabilistic range query）を考える．確率的範囲問合せはデータベースから，問合せオブジェクトから指定した範囲以内に，確率閾値以上の確率で存在するオブジェクトを返す．問合せオブジェクトについては不確実性のない点オブジェクトとガウス分布で表現された不確実なオブジェクトの二通りを想定する．

ガウス分布の性質を分析し，有効なフィルタリング手法を提案した．提案したフィルタリング手法は，確率計算によるコストのかかる検証が必要なオブジェクトの数を大いに減らすことができる．それにより，不必要な計算を避けてコストを節約する．さ

らに，効率的な問合せ処理をサポートするために，新たな索引付け手法を提案し，見込みがないオブジェクトを一つずつではなくグループごとにフィルタリングすることができる．既存の R-木手法を拡張した索引付け手法を開発し，ガウス分布に関する分析に基づいて改善した．提案した索引付け手法は，オブジェクトを効果的に構造化し，問合せ処理の性能を大幅に向上させることができる．実データによる大規模な実験で提案手法の効率性と有効性を実証した．

第二に，最近傍検索（nearest neighbor search）について考える．位置情報に対する最も一般的な問合せとして，距離に基づく最近傍検索には多数の領域で様々なアプリケーションがある．この検索は与えられた問合せ点から最も近いオブジェクトを探す．最近傍検索を，伝統的な位置情報から不確実な位置情報まで拡張する取り組みが様々に行われている．その一例は期待距離（expected distance）である．この距離は不確実な位置情報に対する距離である．この動向に従い，本研究ではガウス分布で不確実な位置を表現し，期待距離でガウスオブジェクトと問合せ点との近さを評価する．以上の設定で，ガウスオブジェクトに対する k-期待最近傍検索（k-expected nearest neighbor search）を考える．結果オブジェクトは問合せ点との期待距離が最小であるトップ k 個のオブジェクトである．

ガウス分布における期待距離の性質を数学的に分析し，この距離の下限と上限を導き出した．この分析に基づいて，効率的に問合せを処理できる三つの新たな手法を提案した．提案手法は実際の期待距離を計算せずに，下限距離が候補オブジェクトの上限距離または期待距離より大きい見込みがないオブジェクトを枝刈りすることができる．候補オブジェクトだけに対して正確な期待距離を計算し，最後に最小であるトップ k 個のオブジェクトを返す．さらに性能を高めるために，R-木を用いてオブジェクトとそれらの下限距離と上限距離に対して索引付けを行う．提案手法は有効的にコストのかかる正確な距離計算の数を減らすことができる．提案手法の効率性と有効性を大規模な実験により検証した．

最後に，類似検索（similarity search）について考える．類似検索はマルチメディアデータベースや，データマイニング，バイオインフォマティクスなどの実世界アプリケーションにおいて不可欠なタスクである．本研究では，ガウス分布で表現された不確実なデータに対する類似検索について研究を行った．問合せオブジェクトもガウス分布により表現する．カルバック・ライブラー情報量（Kullback-Leibler divergence, KL 情報量）を用いて二つのガウス分布の類似度を評価し，データベースから与えられた問合せガウス分布と類似しているトップ k 個のガウス分布を検索する．特に，次元間に相関がなく分散共分散行列が対角である相関なしのガウス分布について考える．

効率的な問合せ処理のために，ランク集約（rank aggregation）とスカイライン問合せ（skyline query）の考え方を用いた新たな二種類の手法を提案した．第一の手法は，データベースの中のすべてのオブジェクトに対して事前に各属性においてソートし，各ソートしたリストからの候補オブジェクトを統合して結果オブジェクトを計算する．第二の手法は，問題を動的スカイライン問合せ（dynamic skyline queries）の計算に変換する．スカイライン問合せ処理のための分枝限定スカイライン（branch-and-bound skyline, BBS）アルゴリズムを拡張・修正し，新たなアルゴリズムを提案した．大規模な実験による性能評価で提案手法の効率性と有効性を示した．

全体として，本論文では，ガウス分布に基づく確率的データの管理に対する包括的な見解を示した．本論文の貢献は多面的であり，さらに時間につれて拡張できると考えている．まず，ガウスオブジェクトに対する確率的範囲問合せ，最近傍検索，および類似検索に対して示した数学的な分析は，既存の関連アプリケーションだけではなく，ガウス分布で表現されたデータに関するほかのアプリケーションにおける研究にも潜在的に役に立つと考えられる．加えて，効率的な問合せ処理のために提案した重要なアルゴリズムと索引構造は，実世界におけるユーザの体験を向上できるのみならず，他の研究問題を解決する際にも有益な洞察と参考となるアイデアを提供することができる．最後に，提案手法の性能における大規模な実験評価を行ったことも本論文の貢献である．

学位関係