平成29年度 情報工学コース卒業研究報告要旨

石川 研究室	氏　　名	安田　健人
卒業研究題目	配列DBMSにおける 空間スキャン統計量の計算手法に関する研究	

近年，計算機の性能向上により，膨大な量のデータが利用できるようになっており，その活用が課題となっている．特に，統計学やパターン認識などの技術を用いることで大規模データから必要な情報を取り出すデータマイニングは大きな注目を浴びている．データマイニングの一種である空間データマイニングの方法は多く提案されているが，その中でも空間的に過密な領域（ホットスポット）を検出する方法は盛んに研究されている．そこで，本研究では空間的に過密な領域を検出する代表的な方法である空間スキャン統計量（spatial scan statistic）に着目する．空間スキャン統計量に関する研究は疫学や生物学への応用や計算の高速化などを中心に多くの研究が行われている．例えば，空間スキャン統計量を計算することにより，病気が流行している地域を検出する研究があり，病気の感染拡大防止に役立っている．しかし，既存研究では，主記憶上での処理が想定されているためスケーラビリティがないことが課題として挙げられる．データが大規模になるほどスケーラビリティの問題は重要になる．

そこで本研究では，近年開発が進んでいる多次元配列形式の大規模データ分析に特化した配列指向DBMS（array-oriented database management system）に注目する．配列指向DBMSは，時空間データや観測データ，シミュレーションデータといった科学データを扱うことを得意としており，配列形式でデータを格納し処理する．配列指向DBMSの1つであるSciDBでは非共有アーキテクチャ（shared nothing architecture）に基づく並列分散処理や，チャンクと呼ばれるデータ構造が取り入れられているため，スケーラビリティが確保され，効率的な処理ができる．

本研究ではSciDBを活用し，空間スキャン統計量を計算する手法について検討を行う．具体的には，$N \times N$の正方形グリッドGに集約されたデータが，SciDB上の配列に格納されている状況を考え，SciDBで提供されている様々な演算を組み合わせて空間スキャン統計量の計算を行うアルゴリズムを実装する．計算した空間スキャン統計量を用いて，空間的に過密な領域を検出するアルゴリズムについても述べる．また，素朴に実装したアルゴリズムの欠点を考察し，その欠点を改善したアルゴリズムを提案する．さらに，素朴なアルゴリズムと改善したアルゴリズムに関して全体の実行時間とアルゴリズムで用いる各演算の実行時間を測定する実験を行い，改善によって実行時間が短縮することを確認する．最後に実験結果を踏まえてアルゴリズムのさらなる高速化に向けた考察を行い，今後の課題について述べる．