論文 / 著書情報
Article / Book Information

題目(和文)	スケーラブルなデータストアにおけるクエリ適応的な多次元索引に関する研究
Title(English)	Query-Adaptive Multidimensional Indexing on Scalable Data Stores
著者(和文)	西村祥治
Author(English)	Shoji Nishimura
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第10874号, 授与年月日:2018年3月26日, 学位の種別:課程博士, 審査員:横田 治夫,宮崎 純,権藤 克彦,吉瀬 謙二,金子 晴彦
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第10874号, Conferred date:2018/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	審査の要旨
Type(English)	Exam Summary

（博士課程）

論文審査の要旨及び審査員

報告番号	甲第　　号		学位申請者氏名		西村　祥治	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	横田　治夫	教授	審査員	金子　晴彦	准教授
	審査員	宮崎　純	教授			
		権藤　克彦	教授			
		吉瀬　謙二	准教授			

論文審査の要旨（2000 字程度）

本論文は「Query-Adaptive Multidimensional Indexing on Scalable Data Stores（スケーラブルなデータストアにおけるクエリ適応的な多次元索引に関する研究）」と題し、優先度の高いクエリに対して実行性能を最適化する多次元索引をスケーラブルなデータストア上で構成する方法およびその効果を論じるもので、英文７章よりなっている。

第 1 章「Introduction」では、大規模なデータに対する分析を実現する場合の、データ管理システムのスケーラビリティと分析処理の高速性の重要性について述べている。近年スケーラブルなデータ管理が実用化されるようになってきているが、データ分析の基本処理である多次元クエリは十分高速化されておらず、また、その高速化には理論的限界があることも知られている。本論文ではクエリ適応性に着目することで解決を図るとし、提案するアプローチとその貢献を概説し、論文の構成を述べている。

第 2 章「Preliminaries」では、本論文の前提知識として、範囲分散型データストア、多次元クエリ、多次元索引および空間充填曲線について概説している。多次元索引により多次元クエリを効率化できる一方で、そのデータ管理モデルは範囲分散型データストアのものとはギャップがあること、さらに、データストアや多次元クエリの性能上のボトルネックが発生することについて論じている。

第 3 章「Related Work」では、多次元クエリの高速化手法として、本論文との関連性が高い多次元索引技術および Data Skipping 技術に関する従来研究を紹介している。既存の多次元索引技術は一般的な場合を対象としており、優先度の高いクエリに適応して性能を最大化する観点が不足している点、また、既存の Data Skipping 技術では多次元クエリに対しては偽陽性率が高く有効に機能していない点を論じている。

第 4 章「MD-HBase: Multidimensional Indexing over Scalable Data Store」では、範囲分散型データストア上に、多次元索引として多用される kd-tree と同等の多次元索引層を、空間充填曲線を適用することで構成する方法を提案している。さらに、その多次元索引層における効率的な多次元範囲検索及び近傍検索アルゴリズムも提案している。提案手法を代表的な範囲分散型データストアである HBase 上に実装し、最大 16 ノードのクラスタ上で評価を行っている。その結果、挿入性能は元の HBase と同等のスループットを達成していること、多次元範囲検索は従来手法より 5〜13 倍に高速化できること、上位 100 件以下の近傍検索が 1.5 秒で達成できることを示している。

第 5 章「QUILTS: Query-Adaptive Data Partitioning」では、4 章の提案手法を拡張し、優先度の高いクエリパターンに対して、そのクエリ性能を最大化するデータ分割手法を提案している。クエリ実行時のページアクセス数に着目し、対象とするクエリパターンに対して、ページアクセス数を最小化するようなデータ分割を実現する空間充填曲線について論じている。そのような空間充填曲線が持つべき性質がデータ分布密度により変化することを明らかにするとともに、それを評価するモデルを提案している。さらに、複合索引として知られている C-カーブと空間索引で広く用いられている Z-カーブを一般化したビット混合曲線族を提案し、その曲線族の中から優先度の高いクエリパターンに適合した空間充填曲線を設計する方法を提案している。実応用である売上分析および地理情報分析を題材に、実データでの性能評価を行い、提案手法で設計した空間充填曲線を用いることで、アクセスページ数を最大で 6 分の 1 に削減でき、また、実行時間をアクセスページ数に比例して短縮できることを示している。さらに、実行時の並列度を上げることでアクセスページ数を超えない並列度で性能が改善できること、および相反するクエリパターンに対する空間充填曲線の設計上の制約を緩和できることを示している。

第 6 章「Multidimensional Range Filters: Compact Index for Data Skipping」では、5 章の空間充填曲線の設計手法を応用し、空間効率の良い多次元範囲フィルタを提案している。このフィルタは与

えられた多次元範囲におけるデータの存在性を返すコンパクトなデータ構造で、検索範囲内にデータが不在の場合、ページアクセスを回避することが可能になる。従来手法では、多次元範囲のデータ存在性を検査する場合、偽陽性率が高くなることが知られている。本論文では、偽陽性の要因分析を基に、対象とするクエリパターンの偽陽性率を低減する空間充填曲線の設計手法、および疎領域のデータ管理手法を提案している。評価実験では、実応用である売上分析と地理情報分析を対象に評価し、フィルタのサイズを元のデータの１～10%に抑えられること、従来手法では 80%あった偽陽性率を最大で 10%まで改善できることを示している。
第７章「Conclusion」では、本論文で提案する内容とその効果をまとめ、発展性について論じている。
以上を要するに、本論文は今後ますます重要度が高まる大規模データ分析の基本処理である多次元クエリを高速化する多次元索引に関して、その根幹をなす独自の複数の手法を提案し、実応用を想定した評価でその効果を示したもので、その適用性、有用性、発展性が高く、その成果は工学上貢献するところが大きい。よって我々は、本論文が博士（工学）の学位論文として十分価値があるものと認める。