氏　　名	Hong Van Le
学位(専攻分野)	博士 (情報学)
学 位 記 番 号	総研大甲第 2078 号
学位授与の日付	平成 31年 3 月22日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Distributed Spatiotemporal Indexes for Querying and Mining in Key-Value Stores
論 文 審 査 委 員	主　査　教授　高須 淳宏 教授　大山　敬三 准教授　片山　紀生 准教授　相原　健郎 教授　原　隆浩　大阪大学大学院 情報科学研究科

(Form 3)

Summary of Doctoral Thesis

Name in full: Hong Van Le

Title: Distributed Spatiotemporal Indexes for Querying and Mining in Key-Value Stores

The recent ubiquity of sensors and GPS-enabled devices has resulted in an explosion of spatiotemporal data generated from probe cars, traffic sensors, and smartphones. To benefit from such data, applications need data storage that can handle the massive volume of data and support high-computational queries. Although key-value store databases (KVSs) efficiently handle large-scale data, they are not equipped with effective functions for supporting geographical data. To solve this problem, we present G$^+$-HBase, a high-performance spatiotemporal database based on HBase, a standard KVS.

First, we present G-HBase, a geographical database based on HBase. To index geographic data, we use Geohash as the rowkey in KVSs. Then, we propose a novel partitioning method, namely binary Geohash rectangle partitioning, to support spatial queries. Our extensive experiments on real datasets have demonstrated improved performance with k nearest neighbors and range query in G-HBase when compared with SpatialHadoop, a state-of-the-art framework with native support for spatial data. We also observed that performance of the spatial join query in G-HBase was on par with SpatialHadoop and outperforms SJMR algorithm in HBase.

Second, we extend G-HBase to G$^+$-HBase to support spatiotemporal data in an intelligent transportation system. In the spatiotemporal index, we adopted STCode, a longitude, latitude, and time-encoding algorithm, to build an index on top of HBase. Our proposed index structure allows continuous updates of objects and provides an efficient prefix filter for supporting spatiotemporal data retrieval. Experimental results demonstrate the high performance of spatiotemporal queries with response time meeting the requirements of real-time query-processing systems.

Finally, we further extend G$^+$-HBase to deal with user-generated geo-tagged social data. We study an efficient multidimensional index structure and parallel processing approaches for the top-k frequent spatiotemporal terms query, a basic analytic query on geo-tagged social data. Given a spatiotemporal range, the query aggregates frequencies of terms among the social posts in that range to find the most frequent terms. In order to reduce storage for indexing and to improve the query performance, we propose a distributed index structure that transforms spatiotemporal coordinates into unique codes

to generate rowkeys in KVSs and balances the data distribution across clusters. Then, we utilize data localization by calculating sorted term lists (STLs) inside storage servers in parallel. To reduce input/output between storage servers and the client, we theoretically estimate the necessary length of STLs to calculate top k frequent terms and send only a part of STLs to the client. From several experiments on real datasets, we observed lower space requirements but better query performance of our approach when compared with baseline approaches.

Form8・Separate Sheet
(様式８・別紙)

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏　　名　Hong Van Le

Title
論文題目　Distributed Spatiotemporal Indexes for Querying and Mining in Key-Value Stores

出願者は、時空間データを効率的に検索・分析するための分散データ処理システムの研究を行い、その成果を博士論文としてまとめた。博士論文における研究は、キーバリュー型の分散データ管理システム上で大規模時空間データを処理するための索引構造と検索・分析アルゴリズムを提案し、その処理性能が既存手法と比較し大きな性能向上を図れることを実験的に示している。

博士論文は７章で構成され英語で書かれている。第１章では、大規模時空間データが近年急速に収集されている背景について説明したのち、その有効利用のために効率的処理技術が必要であることを論じている。続いて、分散データ管理システムによる時空間データ管理の有効性を論じ、本博士論文の主な貢献を示している。第２章では、本研究の基盤となる２つの技術について概説している。まず、本研究で用いたキーバリュー型データ管理システム HBase の概要を述べ、続いて、時空間データの効率的な処理に必要な多次元索引の研究を整理している。第３章では、分散時空間データベースの研究をサーベイし、代表的なシステムの特徴をまとめるとともに、本研究で提案した時空間データベース G+-HBase の優位性を論じている。続く３つの章で本論文の主たる貢献を述べている。第４章では本研究で提案した空間データベースを詳述している。このシステムは、(1)平面をデータベース中のデータの稠密度に応じて分割し、(2)空間充填曲線の１つである Z-order 曲線を用いて各部分空間をハッシュ値に変換し、(3)キーバリュー型システムのキー値として用いるところに特徴がある。この機能を実現するため、効率的なデータ検索に加えデータ更新による空間の再分割のための索引構造である BGRP 木(Binary Geohash Rectangle Partition Tree)を提案している。実験では、約２千万の測地点からなるタクシー移動履歴データおよび約２０億の位置情報から構成される空間オブジェクトのデータを用いて評価実験を行い既存手法と比較し、数倍から数十倍の処理性能向上を得られることを示している。また、データの特性や問い合わせ処理の種類と処理性能との関係について論じている。第５章では、第４章で述べた空間データの処理手法を時空間データに拡張している。ここでは、平面上の位置を表す２次元データに時間を加えた３次元のキー値をハッシュ値に変換する方法を提案し、分散キーバリュー型システムで分散問い合わせ処理を行うアルゴリズムを提案している。第４章と同様に大規模データを用いた評価実験を行い、その高い処理性能を確認している。第６章は、提案した分散時空間データベースを用いた時空間頻出単語列挙アルゴリズムを提案している。この問題は、指定された時空間に関連づけられた文書に現れる k 個の最頻出単語を抽出するもので、５章で提案した分散時空間検索法に、単語の出現頻度情報を組み合わせる方法を提案している。まず、分散システムの各ノード

で指定された時空間領域に関連づけられた文書の単語出現表を作成し、Top K アルゴリズムを利用し各ノードで作成された単語出現表をマージすることで問題を解いている。第7章では、以上の結果をまとめるとともに今後の課題を示している。

公開論文発表会において、出願者はおよそ 45 分で博士論文の内容を説明し、その後、30 分程度の質疑応答が行われた。審査委員からは、関連研究との相違点、提案手法の特徴、実験手順の詳細、本研究成果が有効なデータの種類や今後改善すべき点等について質問とコメントが寄せられ、出願者は適切に回答した。

質疑応答後に審査委員会を開催し、審査委員で議論を行った。博士論文審査の結果、出願者は情報学分野の十分な知識と研究能力を持つと認められ、また研究内容は学位論文として十分なレベルの新規性、有効性があると認められた。本論文の内容に関し、電子情報通信学会論文誌に1編、査読付き国際会議に4編の論文が採択されている。以上より、審査委員会全員一致で、博士論文として十分な水準の研究であると認め、学位の授与に値すると判断した。