



Title	Research on Accurate and Fast Learned Cardinality Estimation for Spatial Data
Author(s)	Ji, Yuchen
Citation	大阪大学, 2025, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/101768
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

論文内容の要旨

氏名 (JI YUCHEN)	
論文題名	Research on Accurate and Fast Learned Cardinality Estimation for Spatial Data (空間データに対する正確かつ高速学習型カーディナリティ推定に関する研究)

論文内容の要旨

Spatial applications are becoming increasingly important in our daily lives. For instance, people can easily find their way to a railway station in a foreign city with mobile map applications. These applications process a large amount of queries to meet users' needs every day. Fast responses to queries are crucial for providing good user experiences. Therefore, efficient query processing is always in demand for spatial applications. Estimating a query's result size, known as the cardinality estimation, plays a significant role in query scheduling, optimization, and processing. For example, query optimization generates optimized query plans based on the estimated cardinality of subqueries. Accurate and fast cardinality estimation substantially improves query efficiency.

Recently, with the advancement of machine learning technologies, many learned methods have been proposed for cardinality estimation. Generally, they have shown greatly improved performance compared with traditional methods like histograms. However, existing methods cannot keep accurate and fast estimations facing following challenges brought by spatial data.

- (i) Complex data distribution: Spatial data are generally multi-dimensional and tend to have a complex data distribution. Some existing methods adopt complex neural networks to approximate the data distribution, resulting in large inference times.
- (ii) Frequent data updates: Data updates are frequent in spatial applications. It is challenging for learned methods to keep a good accuracy on a frequent-updated data distribution. Designing an efficient updating strategy remains an open question for current learning models.
- (iii) Variable data size: Complex spatial data types like polygons have variable sizes. A polygon consists of a variable number of vertices. Therefore, a set of polygons has a large range of possible data sizes. Existing learned solutions cannot handle input data of variable sizes efficiently.

In this thesis, we focus on fast and accurate learned cardinality estimation for spatial data, and address the above challenges. This thesis consists of five chapters.

We introduce the research background and issues for learned cardinality estimation in Chapter 1.

In Chapter 2, we investigate cardinality estimation in the one-dimensional case. We note that the latest proximity query processing methods benefit from learned models (indexes) and have shown better performances than non-learned indexing approaches. Specifically, we address the one-dimensional cardinality estimation problem and consider light-weight learning methods. We first design a prototype of a learned method for one-dimensional cardinality estimation. Second, we empirically evaluate the learned method together with existing methods. Then, we analyze the strong and weak points of these methods and find that the proposed learned method is promising and has the potential for the spatial data case.

In Chapter 3, we focus on dynamic spatial data and propose SAFE (Sampling-Assisted Fast learned

cardinality Estimator).

Dynamic setting assumes frequent data updates, which brings challenges for the task of cardinality estimation.

The estimation methods are expected to be update-friendly and keep fast estimation.

Facing these challenges, we specifically develop a sampling strategy that uses a quad-tree-based data partitioning and extracts a small subset to enable fast training of cardinality estimation models. In addition, we employ two-tier regression models to approximate the spatial data distribution while achieving accurate and fast cardinality estimation.

Furthermore, we provide an incremental model update strategy to avoid re-training all models from scratch when we receive updates.

We conduct experiments on real and synthetic datasets, and their results demonstrate that SAFE (i) outperforms state-of-the-art cardinality estimation models for spatial data and (ii) efficiently handles data updates while ensuring accurate and low-latency estimation.

In Chapter 4, we present PolyCard, a learned cardinality estimator for intersection queries on spatial polygons.

Polygons are the most representative spatial objects and cardinality estimation for polygons has not been well studied.

The commonly used bounding box approximation is intrinsically inaccurate.

Existing learning methods cannot be applied to polygons directly due to the variable sizes of polygons.

We develop an adaptive sampling method to transform polygon data to a fixed size, which considers the spatial distribution of the coordinates belonging to a polygon, and successfully apply learning techniques to spatial polygons.

We propose a training data generator to provide training data distributed over the data space and covering different cardinalities.

Our experiments on four real-world datasets of millions of polygons demonstrate the efficiency and effectiveness of PolyCard.

In Chapter 5, we conclude this thesis and discuss our future work.

Our proposed methods can achieve fast and accurate cardinality estimation for spatial data and can further accelerate query processing.

論文審査の結果の要旨及び担当者

氏 名 (JI YUCHEN)		
論文審査担当者	(職)	氏 名
	主査 教授	原 隆浩
	副査 教授	鬼塚 真
	副査 教授	中野 珠実
	副査 教授	村田 忠彦
	副査 教授	伊達 進
	副査 准教授	佐々木 勇和
	副査 准教授	高橋 慧智

論文審査の結果の要旨

近年、空間アプリケーションの重要性が増している。モバイル地図等のアプリケーションは、ユーザのニーズに応えるために毎日大量のクエリを処理している。クエリに対する迅速な応答は、良好なユーザ体験を提供するために重要であり、効率的なクエリ処理が常に求められている。クエリの結果サイズを推定するカーディナリティ推定は、クエリのスケジューリング、最適化、および処理において重要な役割を果たしており、正確かつ迅速なカーディナリティ推定は、クエリ効率を大幅に向上させる。

本論文では、空間データに対する高速かつ正確な学習型カーディナリティ推定に焦点を当て、以下の課題に対処する手法を提案している。空間データは多次元で複雑なデータ分布を持ち、頻繁なデータ更新や可変サイズのデータを扱う必要があるため、既存の手法では正確かつ迅速な推定が困難であった。本論文の主要な研究成果を要約すると次の通りである。

- (1) 1次元データに対する範囲検索のカーディナリティ推定における軽量な学習手法を提案した。具体的には、探索用の最新の学習インデックスをカーディナリティ推定に活用し、既存手法と比較評価を行った。その結果、提案手法が空間データにも有望であることが示された。
- (2) 動的な空間データに対する範囲検索のカーディナリティ推定器であるSAFEを提案した。この手法では、四分木を用いたデータ分割とサンプリング戦略を開発し、迅速な学習を可能にした。また、2層回帰モデルを用いて空間データ分布を近似し、正確かつ高速なカーディナリティ推定を実現した。さらに、SAFEはデータ更新に対して高い適応性を持ち、頻繁なデータ更新が行われる環境でも安定した性能を発揮することができることを示した。
- (3) 空間ポリゴンに対するインターフェクションクエリの学習型カーディナリティ推定器であるPolyCardを提案した。ポリゴンデータを固定サイズに変換する適応サンプリング手法を開発し、学習技術を適用した。この手法は、ポリゴンの形状やサイズの多様性を考慮し、効率的なデータ処理を可能にする。また、PolyCardは、ポリゴンデータの特性を活かした学習モデルを構築することで、高い精度と効率を両立させている。

以上のように、本論文は空間データに対する高速かつ正確なカーディナリティ推定を実現するための基礎技術を提供している。特に、動的なデータ更新や可変サイズのデータに対応する手法は新規性が高く、空間アプリケーションのクエリ処理を加速する上で重要な貢献を果たしている。よって本論文は博士（情報科学）の学位論文として価値のあるものと認める。