Title

High-Performance Processing of Continuous Uncertain Data

Author

Thanh Thi Lac Tran, *University of Massachusetts Amherst* Follow

Date of Award

5-2013

Document Type

Open Access Dissertation

Degree Name

Doctor of Philosophy (PhD)

Degree Program

Computer Science

First Advisor

Yanlei Diao

Second Advisor

Jim Kurose

Third Advisor

Anna Liu

Subject Categories

Computer Sciences

Abstract

Uncertain data has arisen in a growing number of applications such as sensor networks, RFID systems, weather radar networks, and digital sky surveys. The fact that the raw data in these applications is often incomplete, imprecise and even misleading has two implications: (i) the raw data is not suitable for direct querying, (ii) feeding the uncertain data into existing systems produces results of unknown quality.

This thesis presents a system for uncertain data processing that has two key functionalities, (i) capturing and transforming raw noisy data to rich queriable tuples that carry attributes needed for query processing with quantified uncertainty, and (ii) performing query processing on such tuples, which captures changes of uncertainty as data goes through various query operators. The proposed system considers data naturally captured by continuous distributions, which is prevalent in sensing and scientific applications.

The first part of the thesis addresses data capture and transformation by proposing a probabilistic modeling and inference approach. Since this task is application-specific and requires domain knowledge, this approach is demonstrated for RFID data from mobile readers. More specifically, the proposed solution involves an inference and cleaning substrate to transform raw RFID data streams to object location tuple streams where locations are inferred from raw noisy data and their uncertain values are captured by probability distributions.

The second, also the main part, of this thesis examines query processing for uncertain data modeled by continuous random variables. The proposed system includes new data models and algorithms for relational processing, with a focus on aggregation and conditioning operations. For operations of high complexity, optimizations including approximations with guaranteed error bounds are considered. Then complex queries involving a mix of operations are addressed by query planning, which given a query, finds an efficient plan that meets user-defined accuracy requirements.

Besides relational processing, this thesis also provides the support for user-defined functions (UDFs) on uncertain data, which aims to compute the output distribution given uncertain input and a black-box UDF. The proposed solution employs a learning-based approach using Gaussian processes to compute approximate output with error bounds, and a suite of optimizations for high performance in online settings such as data stream processing and interactive data analysis.

The techniques proposed in this thesis are thoroughly evaluated using both synthetic data with controlled properties and various real-world datasets from the domains of severe weather monitoring, object tracking using RFID readers, and computational astrophysics. The experimental results show that these techniques can yield high accuracy, meet stream speeds, and outperform existing techniques such as Monte Carlo sampling for many important workloads

.

Recommended Citation

Tran, Thanh Thi Lac, "High-Performance Processing of Continuous Uncertain Data" (2013). *Open Access Dissertations*. 768.
https://scholarworks.umass.edu/open_access_dissertations/768