REAL-TIME EXPLORATION OF LARGE SPATIOTEMPORAL DATASETS BASED ON ORDER STATISTICS

Original Work by: Cicero Pahins, Nivan Ferreira, and João Comba
Presented By: Vaastav Anand

OUTLINE
- Background
- Motivation
- Quantile Databse Structure
- Example Visualizations
- Evaluation
- Critique

SPATIOTEMPORAL DATASETS

- Datasets generated from measuring a set of values across a set of locations (spatial dimension) across a time range (temporal dimension).

EXISTING SOLUTIONS

- Precomputed indices that store aggregations of a given dataset to solve this problem.
- Gaussian Cubes that support interactive data modelling by describing data distribution using parametric Gaussian distributions.
- Based on robust statistics (mean + covariance).
- Can't assume real-world data has a normal distribution.

OUTLINE
- Background
- Motivation
- Quantile Databse Structure
- Example Visualizations
- Evaluation
- Critique

MEMORY FOOTPRINT ISSUES

- Memory footprint is too high.
- Distributions are approximated using non-robust statistics.
- Queries are slow enough to disallow interactive experience.
- Queries are limited to count queries.

EXAMPLE QUESTIONS

- How likely is a flight operated by Delta Airlines to be delayed more than 10 minutes at JFK airport?
- How unusual are the delays experienced by Delta flights on January 29th, 2017?
- How unusual are the delays experienced by Delta flights on January 29th, 2017?

INDEXING SCHEME

- A novel data structure.
- Encode data distributions based on robust statistics.
- Uses a non-parametric modelling technique called p-digest.
- A novel indexing structure that reduces the large memory footprint.

INDEXING SCHEME

- A novel data structure.
- Encode data distributions based on robust statistics.
- Uses a non-parametric modelling technique called p-digest.
- A novel indexing structure that reduces the large memory footprint.

P-DIGEST DATA SKETCH

- "Data Sketch": Data Structure that can be easily updated with new or modified data and supports a fast set of queries whose results approximate queries on the full dataset.
- An optimized version of p-digest data sketch.
- "Quantile sketch": A sketch that supports queries on quantiles and cdf estimation.
- Summarizes the empirical distribution of an input dataset by a set of weighted values called centroids.

P-DIGEST DATA SKETCH

- "Data Sketch": Data Structure that can be easily updated with new or modified data and supports a fast set of queries whose results approximate queries on the full dataset.
- An optimized version of p-digest data sketch.
- "Quantile sketch": A sketch that supports queries on quantiles and cdf estimation.
- Summarizes the empirical distribution of an input dataset by a set of weighted values called centroids.

QUERY ALGORITHM

- "Quantile sketch": A sketch that supports queries on quantiles and cdf estimation.
- Summarizes the empirical distribution of an input dataset by a set of weighted values called centroids.

DATES_FS

- Datasets generated from measuring a set of values across a set of locations (spatial dimension) across a time range (temporal dimension).

QUERY ALGORITHM

- "Quantile sketch": A sketch that supports queries on quantiles and cdf estimation.
- Summarizes the empirical distribution of an input dataset by a set of weighted values called centroids.
QUANTILE HEATMAPS
- Instead of using the mean for a given location, use the specified quantile at the location as the aggregate measure
- Quantiles are not sensitive to outliers whereas mean is.
- Powered by QDS's quantile queries

CDF HEATMAPS
- Instead of showing mean, they show high likely a distribution in a given location is to be smaller than a certain value
- Powered by QDS's cdfquery

READABLE UNCERTAINTY VISUALIZATIONS
- Interpreting uncertainty visualizations is not easy.
- Can have high cognitive load
- May researchers have to switch the cognitive load
- But existing interactive techniques are not efficient
- QDS’s cdfquery can allow for this interactivity by computing the result for each box plot.

OUTLIER EXPLORATION
- Finding outliers requires users to inspect a large number of data slices over time and space.
- QDS can retrieve approx. distributions over an arbitrary portion of the data very quickly.
- Authors define an outlierness measure supported by QDS’s cdfquery

CONCLUSION
- Presents QDS: a fast in-memory data structure
- Supports uncertainty exploration + data distribution estimation
- Better Memory Footprint
- Comparably better than HashedCubes
- Better for some datasets, Worse for other datasets
- Large Static Spatiotemporal datasets
- MonetDB, SQLite, PostgreSQL don't provide effective implementation for spatial filtering with temporal and categorical constants
- Inefficient interactive techniques are not efficient
- QDS can retrieve approx. distributions over an arbitrary portion of the data very quickly
- Authors define an outlierness measure supported by QDS’s cdfquery

OUTLINE
- Background
- Motivation
- Quantile Datacube Structure
- Example Visualizations
- Evaluation
- Critique

STRENGTHS
- Thorough performance eval
- Good comparisons vs believable baselines
- Built on/Promotes usage of Robust Statistics
- Allows for exploration of uncertainty in datasets.

WEAKNESSES
- Lack of Validation: No User Study
- One of the goals was more interactivity. Can’t validate w/o user study.